Constrained Likelihood Inference in Instrumental Variable Regression with Invalid Instruments and Its Application to GWAS Summary Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Haoran Xue

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advised by Dr. Xiaotong Shen and Dr. Wei Pan

May, 2021

Acknowledgements

Looking back at my past few years in Minnesota, I feel extremely fortunate to be supported by people around me. Foremost, I would like to express my deepest gratitude to my advisors Dr. Xiaotong Shen and Dr. Wei Pan. Dr. Shen has always been supportive in my research and life, his guidance benefited me a lot, his encouragement and optimism helped me through hard times. Dr. Pan has taught me what a qualified researcher should be, and his support in my personal development really means a lot to me, especially during the past year. Their mentorship enables me to finish my graduate study.

I would thank my other committee members, Dr. Adam Rothman and Dr. Lan Liu, for their valuable feedbacks for my dissertation. I would thank other professors in Minnesota who have taught me or helped me, and thank the School of Statistics for offering me the opportunity to study here. Also, I want to thank my friends in Minnesota and other places.

I would like to thank my parents, for bringing me to the world, for their care and love.

Dedication

To my father Wei Xue and my mother Fengmei Sun.

Abstract

There has been increasing interest in instrumental variables regression for causal inference. In genetics, transcriptome-wide association studies (TWAS), also known as PrediXcan, have recently emerged as a widely applied tool to discover causal/target genes by integrating an outcome GWAS dataset with another gene expression/ transcriptome GWAS (called eQTL) dataset; they can not only boost statistical power but also offer biological insights by identifying (putative) causal genes for a GWAS trait, e.g. low-density lipoprotein cholesterol (LDL). Statistically TWAS apply (two-sample) two-stage least squares (2SLS) with multiple correlated SNPs as instrumental variables (IVs) to predict/impute gene expression, in contrast to typical (two-sample) Mendelian randomization (MR) approaches using independent SNPs as IVs, which are expected to be lower-powered. However, some of the SNPs used may not be valid IVs as a result of their (horizontal) pleiotropic/direct effects on the trait not mediated through the gene of interest, leading to false conclusions by TWAS (or MR). We propose a general inferential method for possibly high-dimensional data to account for confounding and invalid IVs while selecting valid IVs simultaneously via two-stage constrained maximum likelihood; we develop a theory for the likelihood method subject to a truncated L_1-constraint approximating the L_0-constraint for asymptotically valid and efficient statistical inference on causal effects. We demonstrate both theoretically and numerically the superior performance of the proposed method over the standard 2SLS/TWAS and other methods. We apply the methods to identify causal genes for LDL by integrating GWAS summary data with eQTL data.

Contents

Acknowledgements **i**

Dedication **ii**

Abstract **iii**

List of Tables **vi**

List of Figures **vii**

1 Introduction **1**

2 Methods **6**

2.1 Model 6

2.2 One-Sample Case 8

2.2.1 Oracle Estimator 8

2.2.2 New Method: Two-stage Constrained Maximum Likelihood . . 10

2.2.3 Tuning parameter selection with BIC 13

2.3 Two-Sample Case 14

2.3.1 Oracle Estimator 14

2.3.2 New Method: Two-stage Constrained Maximum Likelihood . . 16

2.3.3 Tuning parameter selection with BIC 17

2.4 Computation 17

2.5 Extension to GWAS Summary Data 19

3 Simulations **22**

3.1 Simulation 1: One-Sample Case 22

3.2 Simulation 2: Two-Sample Case 25

3.3 Other Simulations 29

4 Real Data Example **30**

5 Conclusion and Discussion **33**

References **33**

A Proofs **40**

B Consistent Estimation of the Asymptotic Variances **71**

B.1 One-Sample Case 71

B.1.1 Estimating v with the Oracle Estimators 71

B.1.2 Estimating v with the 2ScML Estimators 73

B.2 Two-Sample Case 74

B.2.1 Estimating v with the Oracle Estimators 74

B.2.2 Estimating v with the 2ScML estimators 75

C More Simulation Results **76**

C.1 Full Simulation 1 Results 76

C.2 Simulation 3: Setups Mimicking Real Data 80

D Real Data Example: More Results **87**

References for Appendix **93**

List of Tables

3.1 The means ($\bar{\hat{\beta}}$ and $\bar{SE}(\hat{\beta})$) of the estimates $\hat{\beta}$ and their standard errors, and the standard deviations (sd) of $\hat{\beta}$ for MR-Egger, 2ScML and the oracle methods in Simulation 2 with $\sigma_\phi = 0$ (i.e. with IV Assumption (B) and thus the InSIDE assumption holding). 28

3.2 The means ($\bar{\hat{\beta}}$ and $\bar{SE}(\hat{\beta})$) of the estimates $\hat{\beta}$ and their standard errors, and the standard deviations (sd) of $\hat{\beta}$ for MR-Egger, 2ScML and the oracle methods in Simulation 2 with $\sigma_\phi = 0.1$ (i.e. with IV Assumption (B) and thus the InSIDE assumption violated). 29

4.1 The eight genes each with a score of 4 or 5 and identified by TWAS or 2ScML to be associated with LDL. 32

C.1 Empirical type I error rates (for $\beta^0 = 0$) and power (for $\beta^0 \neq 0$) of the methods in Simulation 1. 79

C.2 Ten relevant IVs/SNPs on chromosome 19 used to generate exposure D, and their correlations with invalid IVs/SNPs rs8100875, noRSname and rs2288918. 81

D.1 The 32 significant genes associated with LDL identified by TWAS or/and 2ScML with their literature search support scores (Score) and corresponding references (Refs). The p-values less than the significance cut-off 0.05/4580 are marked red. 88

List of Figures

2.1 The true causal model for (2.1). Directed edges represent direct effects; both $\boldsymbol{\gamma}_A^0$ and $\boldsymbol{\alpha}_B^0$ are non-zero; depending on whether $\beta^0 \neq 0$ or not, D has or does not have a causal effect on Y. 7

3.1 Empirical Type-I error rates (left panel) and power (right panel) for Simulation 1 (Setup 1) at the nominal level 0.05. 23

3.2 Simulation 2: empirical Type I error rates (for $\beta^0 = 0$) and power (for $\beta^0 \neq 0$) when IV Assumption (B) was not (upper panel) or was (lower) violated. 27

4.1 Comparison of the p-values in the $-\log_{10}$ scale of the 32 significant genes for LDL. 32

C.1 Empirical Type-I Error Rates of Setup 1 and Setup 2: the x-axis shows the sample size, while y-axis shows the empirical Type-I Error rates based on 1000 simulations; the horizontal dashed line represents the nominal level 0.05. 77

C.2 Empirical Power Rates of Setup 1 and Setup 2: the x-axis shows the causal effect size β^0, while y-axis shows the empirical power rate based on 1000 simulations; the horizontal dashed line represents the nominal level 0.05. 78

C.3 Generating simulated data for gene GEMIN7 in two setups. 83

C.4 Simulation results for gene GEMIN7 with 3 IVs being invalid. 84
C.5 Simulation results for gene GEMIN7 with all IVs being valid. 85

Chapter 1

Introduction

Transcriptome-wide association studies (TWAS), as implemented in PrediXcan [12] and TWAS [15], were recently proposed to boost statistical power and enhance interpretation. It was motivated by one key hypothesis that many genetic variants influence complex traits through transcriptional regulation. They have soon become popular with applications to common diseases like type 2 diabetes, schizophrenia, and cancer, convincingly showing the power of integrating genome-wide association studies (GWAS) and expression quantitative trait locus (eQTL) data to gain biological insights. Specifically, TWAS implicate (putative) causal genes of a GWAS trait, overcoming a severe limitation of GWAS in a lack of biological insights from GWAS discoveries of trait-associated genetic variants. Statistically TWAS apply the standard (two-sample) two-stage least squares (2SLS) in the framework of instrumental variable (IV) regression for causal inference. IV regression is a general and powerful tool for estimating and drawing inference about a causal effect from the exposure to an outcome in the presence of unmeasured confounding. A valid IV must satisfy three assumptions:

(A). Relevance: it is associated with the exposure;

(B). Exchangeability: it is not associated with unmeasured confounders;

(C). Exclusion restriction: it is not associated with the outcome conditional on the exposure.

Given valid IVs, 2SLS makes a correct inference about the causal effect; yet it may break down and give erroneous results in the presence of invalid IVs. Assumption (A) ensures the inclusion of *relevant* IVs, which is more straightforward and typically possibly more conservatively handled by using a stringent significance cut-off. In contrast, testing assumptions (B) or (C) is more challenging; between (B) and (C), the former is even harder (due to the *hidden* confounding) while the existing literature (especially concerning MR) is more focused on (C). As to be discussed, the proposed method can deal with the violation of all three assumptions. Kang et al. [17] proposed a lasso-type method called *sisVIVE* for estimating causal effect with some invalid IVs but did not address the problem of inference. Lin et al. [18] proposed a two-stage regularization method to select optimal instruments and jointly estimate the effects of multiple exposures on the outcome, but they did not allow invalid IVs in stage 2 and did not consider the problem of inference either. Windmeijer et al. [36] proposed a two-stage method to make inference about the causal effect in the low-dimensional setting with a fixed number of instruments. Because of the median estimator used in the first stage, their method requires the "Majority Condition", that is, more than 50% of the instruments are valid. When the "Majority Condition" fails but a weaker "Plurality Condition" holds, *Two-Stage Hard Threshold* (TSHT) by Guo et al. [14] can handle inference in the low-dimensional situation with the oracle property; for the high-dimensional case, TSHT can consistently estimate the set of valid instruments and then make correct inference but without the oracle property. Importantly, the aforementioned works all deal with the one-sample case, in which the data used for the two-stage modeling are collected from the same sample of individuals. In contrast, the two-sample case has dominated recent genetic applications in TWAS and MR as to be discussed later, in which the exposure and the outcome data are from two

independent samples.

We propose a *Two-Stage Constrained Maximum Likelihood* (2ScML) method to make inference on causal effects in the framework of instrumental variables regression as 2SLS. First, we aim to tackle the problem in a more general setting than that of many other methods. In particular, we allow high-dimensional data in the presence of invalid IVs with all three IV assumptions violated. Compared to some existing methods with two different initial and final estimators, we propose a unified constrained regression approach for simultaneous variable selection, accounting for invalid IVs and drawing inference at the same time. Second, in contrast to TSHT, in the high-dimensional setting, our method can consistently identify the set of invalid instruments in the second stage and thus has the oracle property. Third, we extend our method to GWAS summary data, largely broadening its applications to genetics where individual-level data from large-scale GWAS are often unavailable. In particular, the two-sample design has dominated recent genetic studies with easy and wide applications to two independent GWAS summary datasets on exposure and an outcome respectively. For this purpose, we develop our method for the two-sample case, in addition to the one-sample case. We also propose BIC for consistent model selection; it is applicable with either GWAS individual-level data or summary data. We are not aware of any other existing methods with all the above features of our proposed method.

The proposed method is especially suitable for applications to TWAS to identify causal genes or other molecular/imaging/clinical endophenotypes by integrating GWAS with other eQTL/xQTL data [12, 15, 44, 45, 38, 39, 30, 6, 16]. In these applications, multiple correlated SNPs (so-called cis-SNPs) near a gene are used as IVs to impute or predict the gene's expression level (or another endophenotype) to infer whether the gene's expression (or another risk factor) has a causal effect on a trait, say low-density lipoprotein cholesterol (LDL). However, due to strong modeling

assumptions on valid IVs that may be violated frequently in practice, cautions have to be taken about the conclusions from the standard TWAS. For example, it is known that TWAS tends to identify multiple genes per locus, most of which are likely false positives due to confounding caused by linkage disequilibrium (LD) among nearby SNPs [19, 34, 37]. In particular, due to confounding through LD between an eQTL (i.e. an SNP causal to a gene's expression) and a true causal SNP to a GWAS trait, a target gene identified by TWAS (or MR) may be only marginally *associated with, but not causal to,* the GWAS trait, similar to that of a significant tagging SNP in GWAS may not be causal. Furthermore, due to widespread (horizontal) pleiotropy [33], some SNPs used in TWAS may not be valid IVs, again leading to violations of a critical assumption in TWAS/2SLS [3, 5]. As alternatives to TWAS, another class of popular IV analysis using (often independent) SNPs as IVs is (two-sample) Mendelian randomization (MR) [8, 9, 10]. In these applications, due to often a small sample size of an eQTL study (i.e. Stage 1 in 2SLS), it would be low powered to apply a single SNP/IV-based method as in MR, as implemented in SMR and GSMR for the same purpose [44, 45]; instead, it would be more powerful and thus more desirable to apply a method with multiple SNPs used to predict the gene's expression level (or another exposure/trait in the first stage). Furthermore, an MR method, requiring every single SNP to be a valid IV and thus to be associated with the exposure (i.e. a gene's expression level), in contrast to that the collection of the multiple SNPs to be associated with the exposure in TWAS, will be more likely to encounter the bias problem of weak IVs. For these two reasons, we will focus on TWAS, not MR or its extensions (e.g. some popular ones reviewed in [29] or new ones described in [31, 42]), though we will briefly compare with a new and perhaps the most popular MR approach called Egger regression, which is robust in the presence of invalid IV with IV Assumption (C) violated [4]. However, if IV Assumption (B) is violated, MR-Egger regression would break down, as to be shown in our simulation study; besides, we will

show much higher statistical efficiency/power of our new method over MR-Egger regression. Finally, we also point out that the proposed method, as a general extension to 2SLS, can be also applied to other problems, e.g. inferring causal relationships between pairs of traits based on pairs of GWAS.

Chapter 2

Methods

2.1 Model

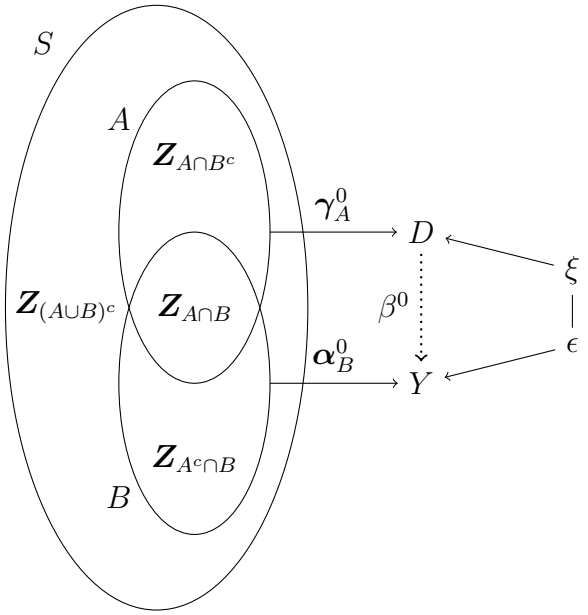
We denote an exposure as $D \in \mathbb{R}$, an outcome of interest as $Y \in \mathbb{R}$, p IVs (such as SNPs) as $\boldsymbol{Z} \in \mathbb{R}^p$. In the following, for a subset $G \subseteq S = \{1, 2, \cdots, p\}$ and vector $\boldsymbol{V} \in \mathbb{R}^p$, $\boldsymbol{V}_G$ is the corresponding sub-vector of $\boldsymbol{V}$. Suppose we have n i.i.d. samples $\{(Y_i, D_i, \boldsymbol{Z}_i)|i = 1, \cdots, n\}$. Corresponding to the true causal model in Figure 2.1, our statistical models for the exposure and the outcome are

$$\begin{aligned} D_i &= \boldsymbol{Z}_i^T \boldsymbol{\gamma}^0 + \xi_i, \\ Y_i &= \beta^0 \cdot D_i + \boldsymbol{Z}_i^T \boldsymbol{\alpha}^0 + \epsilon_i. \end{aligned} \tag{2.1}$$

Here $\begin{pmatrix} \epsilon_i \\ \xi_i \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix})$ are the error terms independent of instruments $\boldsymbol{Z}_i$; $\boldsymbol{\gamma}^0 \in \mathbb{R}^p$ are the true effects of the IVs on the exposure, and for some $A \subseteq S$, $\boldsymbol{\gamma}_A^0 \neq 0$, $\boldsymbol{\gamma}_{A^c}^0 = 0$; $\beta^0 \in \mathbb{R}$ is the parameter of interest, representing the causal effect of D on Y; $\boldsymbol{\alpha}^0 \in \mathbb{R}^p$ are the direct effects of the IVs on Y, and for some $B \subseteq S$, $\boldsymbol{\alpha}_B^0 \neq 0$, $\boldsymbol{\alpha}_{B^c}^0 = 0$, and $|B| = p_0$. Note that, if B is not empty, it explicitly accounts for the violation of IV assumptions (B) or/and (C), a main problem to be addressed here.

Subsequently, we assume that by default the above two-stage linear models in

Figure 2.1: The true causal model for (2.1). Directed edges represent direct effects; both $\boldsymbol{\gamma}_A^0$ and $\boldsymbol{\alpha}_B^0$ are non-zero; depending on whether $\beta^0 \neq 0$ or not, D has or does not have a causal effect on Y.



(2.1) hold for all our later theorems. We also note that after centering all variables at the sample mean 0, we do not have the intercepts in the two models in (2.1).

One primary aim is to infer β^0 or the causal effect of the exposure D_i on the outcome Y_i. Note that in general D_i and ϵ_i are not independent due to $\sigma_{12} \neq 0$; for this reason, even under the low-dimensional setting, the ordinary least squares (OLS) gives a biased estimate of β^0 (both in finite samples and asymptotically), and 2SLS in the general framework of instrument variable regression has been proposed for (asymptotically) unbiased inference. The two models in (2.1) correspond to that for the two stages, respectively.

In what is to follow, we use the following notations: $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T \in \mathbb{R}^{n\times 1}$, $\boldsymbol{D} = (D_1, \cdots, D_n)^T \in \mathbb{R}^{n\times 1}$, $\mathbf{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)^T \in \mathbb{R}^{n\times p}$, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \in \mathbb{R}^{n\times 1}$, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)^T \in \mathbb{R}^{n\times 1}$. For any set $G \subseteq S$, we use $\mathbf{Z}_G$ to denote the corresponding

columns of matrix $\mathbf{Z}$. We assume that $\mathbf{Z}_A$ and $\mathbf{Z}_B$ are of full rank in column. For any matrix $\mathbf{X}$, $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix to the column space of $\mathbf{X}$, and $\mathbf{M_X} = \mathbf{I} - \mathbf{P_X}$ is the residual projection matrix. $\mathbf{X}_{ij}$ is the element in the i^{th} row and j^{th} column of $\mathbf{X}$.

The Plurality Condition, as stated in Guo et al. [14], is both sufficient and necessary for parameter identifiability in model (2.1). For completeness we show the theorem; see [14] for a proof.

Theorem 1. ***(Guo et al. 2018)*** *Assume that $E(\boldsymbol{Z}_i\boldsymbol{Z}_i^T)$ is invertible. Then model (2.1) is identifiable if and only if the Plurality Condition holds:*

$$|A \cap B^c| > \max_{c \neq 0} |j \in A : \alpha_j/\gamma_j = c|.$$

Here $A \cap B^c$ is the set of valid IVs satisfying all three IV assumptions. The Plurality Condition is always assumed in the following and we will not state it explicitly again.

2.2 One-Sample Case

2.2.1 Oracle Estimator

Suppose we have a sample of size n containing $\boldsymbol{Z}$, D, and Y from model (2.1). In the ideal (but unrealistic) situation with the two sets A and B known, we define the two-stage oracle estimator

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_A^{or} &= \underset{\boldsymbol{\gamma}_A}{\operatorname{argmin}} ||\boldsymbol{D} - \mathbf{Z}_A\boldsymbol{\gamma}_A||^2, \quad \hat{\boldsymbol{D}} = \mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A^{or}, \\ (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}_B^{or}) &= \underset{\beta, \boldsymbol{\alpha}_B}{\operatorname{argmin}} ||\boldsymbol{Y} - \beta \cdot \hat{\boldsymbol{D}} - \mathbf{Z}_B\boldsymbol{\alpha}_B||^2, \end{aligned} \tag{2.2}$$

where $||\cdot||$ represents the L_2-norm. By setting $\hat{\gamma}_i^{or} = 0$ for $i \notin A$, we can extend $\hat{\boldsymbol{\gamma}}_A^{or}$ to $\hat{\boldsymbol{\gamma}}^{or}$; and by setting $\hat{\alpha}_i^{or} = 0$ for $i \notin B$, we can extend $\hat{\boldsymbol{\alpha}}_B^{or}$ to $\hat{\boldsymbol{\alpha}}^{or}$. The following

assumptions are made to ensure the uniqueness and desired properties of the oracle estimator. Let $\mathbf{X}_0 = (\mathbf{Z}_A\boldsymbol{\gamma}_A^0, \mathbf{Z}_B)$.

Assumption 1. *Assume* $||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0||^2/n \geq d_0$, *where* d_0 *is some positive constant. Assume that as* $n \to \infty$, $\mathbf{Z}_A^T\mathbf{Z}_A/n \to \mathbf{U}_A$ *(with its inverse* $\mathbf{U}_A^{-1}$*);* $\mathbf{Z}_A^T\mathbf{Z}_B/n \to \mathbf{U}_{AB}$*;* $\mathbf{Z}_B^T\mathbf{Z}_B/n \to \mathbf{U}_B$. *Then*

$$\frac{\mathbf{X}_0^T\mathbf{X}_0}{n} = \begin{pmatrix} (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0 & (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_B \\ \mathbf{Z}_B^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0 & \mathbf{Z}_B^T\mathbf{Z}_B \end{pmatrix}/n \to \begin{pmatrix} (\boldsymbol{\gamma}_A^0)^T\mathbf{U}_A\boldsymbol{\gamma}_A^0 & (\boldsymbol{\gamma}_A^0)^T\mathbf{U}_{AB} \\ \mathbf{U}_{AB}^T\boldsymbol{\gamma}_A^0 & \mathbf{U}_B \end{pmatrix} := \boldsymbol{\Sigma}$$

with its inverse $\boldsymbol{\Sigma}^{-1}$, *and*

$$\begin{aligned} \frac{\mathbf{X}_0^T\mathbf{P}_{\mathbf{Z}_A}\mathbf{X}_0}{n} &= \begin{pmatrix} (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0 & (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_B \\ \mathbf{Z}_B^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0 & \mathbf{Z}_B^T\mathbf{P}_{\mathbf{Z}_A}\mathbf{Z}_B \end{pmatrix}/n \\ &\to \begin{pmatrix} (\boldsymbol{\gamma}_A^0)^T\mathbf{U}_A\boldsymbol{\gamma}_A^0 & (\boldsymbol{\gamma}_A^0)^T\mathbf{U}_{AB} \\ \mathbf{U}_{AB}^T\boldsymbol{\gamma}_A^0 & \mathbf{U}_{AB}^T\mathbf{U}_A^{-1}\mathbf{U}_{AB} \end{pmatrix} := \boldsymbol{\Psi}. \end{aligned}$$

Assumption 1 says that $\mathbf{Z}_A\boldsymbol{\gamma}_A^0$, the total effect of instruments $\mathbf{Z}_A$ on exposure $\boldsymbol{D}$, is separated from the column space of instruments $\mathbf{Z}_B$, which have direct effects on outcome $\boldsymbol{Y}$. Moreover, the covariance matrices are required to converge. Theorem 2 describes some statistical properties of the oracle estimator.

Theorem 2. *With Assumption 1 satisfied, the probability of the oracle estimator* $\hat{\beta}^{or}$ *defined in (2.2) being unique converges to 1 as* $n \to \infty$, *and* $\hat{\beta}^{or}$ *is a consistent estimator of the true causal effect* β^0 *with* $\hat{\beta}^{or} \xrightarrow{p} \beta^0$ *as* $n \to \infty$. *Furthermore, we have* $\sqrt{n}(\hat{\beta}^{or} - \beta^0) \xrightarrow{d} N(0, v)$, *with variance* $v = (\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2) \cdot (\boldsymbol{\Sigma}^{-1})_{11} - (2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2) \cdot (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})_{11}$. *Under the null hypothesis* H_0*:* $\beta^0 = 0$, *we have* $v = \sigma_1^2 \cdot (\boldsymbol{\Sigma}^{-1})_{11}$. *Here* $\sigma_1^2, \sigma_2^2, \sigma_{12}$ *are the variances and covariance of the error terms as defined in (2.1).*

A consistent estimator of the asymptotic variance v is

$$\hat{v}^{or} = \hat{v}_1^{or} \cdot (\hat{\mathbf{\Sigma}}^{or})_{11}^{-1} - (\hat{v}_1^{or} - \hat{v}_2^{or}) \cdot \left((\hat{\mathbf{\Sigma}}^{or})^{-1} \hat{\mathbf{\Psi}}^{or} (\hat{\mathbf{\Sigma}}^{or})^{-1} \right)_{11} \xrightarrow{p} v,$$

where $\hat{v}_1^{or}$ and $\hat{v}_2^{or}$ are consistent estimates of $(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)$ and σ_1^2, and $\hat{\mathbf{\Sigma}}^{or}$ and $\hat{\mathbf{\Psi}}^{or}$ are consistent estimates of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, respectively. Details are given in the Appendix.

2.2.2 New Method: Two-stage Constrained Maximum Likelihood

The proposed method consists of two stages as an extension to 2SLS. In the first stage, we solve a constrained maximum likelihood problem to select relevant IVs to satisfy Assumption (A), as the approach of [28] for a general regression model:

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} ||\boldsymbol{D} - \mathbf{Z}\boldsymbol{\gamma}||^2 \text{ subject to } \frac{1}{\tau_1} \sum_{j=1}^{p} \min(|\gamma_j|, \tau_1) \leq K_1, \tag{2.3}$$

where $\frac{1}{\tau_1}\min(|\gamma_j|, \tau_1)$ [27] is the truncated L_1-function for γ_j, which is a continuous surrogate of the L_0 loss $I(\gamma_j \neq 0)$ with $I(\cdot)$ the indicator function. The tuning parameter K_1 can be interpreted as the number of non-zero components of $\boldsymbol{\gamma}^0$, and the constrained problem (2.3) performs a best-subset-like (but computationally much more efficient) search to select K_1 relevant IVs. In practice, we chooseτ_1 to be a small value like 1×10^{-5} and use cross-validation or BIC to estimate the optimal K_1 from a set of candidate integers.

Given $\hat{\boldsymbol{\gamma}}$, we obtain the predicted exposure as $\hat{D}_i = \boldsymbol{Z}_i\hat{\boldsymbol{\gamma}}$; denote $\hat{\boldsymbol{D}} =$

$(\hat{D}_1, \hat{D}_2, \cdots, \hat{D}_n)^T$. Then in the second stage we solve constrained minimization:

$$(\hat{\beta}, \hat{\boldsymbol{\alpha}}) = \underset{\beta, \boldsymbol{\alpha}}{\operatorname{argmin}} ||\boldsymbol{Y} - \beta \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\boldsymbol{\alpha}||^2 \text{ subject to } \frac{1}{\tau_2} \sum_{j=1}^{p} \min(|\alpha_j|, \tau_2) \leq K_2. \quad (2.4)$$

Again, we choose τ_2 to be a small value like 1×10^{-5} and use cross-validation or BIC to estimate the optimal K_2, or the number of invalid IVs. Here we model the direct effects of the IVs explicitly, and use the non-convex constraint to select and thus account for invalid IVs that violate the IV Assumptions (B) and (C).

It is noted that, under the normality assumption, in each stage, the objective function of the proposed method is the squared error loss function as used in 2SLS, though a truncated L_1 constraint (TLC) is imposed to select relevant IVs and invalid IVs respectively in the two stages. We refer to our method as the constrained maximum likelihood in anticipation of its extensions to other parametric models.

Define $\mathbb{A} = \{A_1 \big| A_1 \subseteq S, |A_1| \leq |A|, A_1 \neq A\}$, $\mathbb{B} = \{B_1 \big| B_1 \subseteq S, |B_1| \leq p_0, B_1 \neq B\}$, and $\mathbb{G} = \{B_1 \big| B_1 \subseteq S, |B_1| \leq p_0, B_1 \neq B, A \nsubseteq B_1\}$, and denote $\sigma_M^2 = \max(\sigma_1^2, \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2)$, $\sigma_m^2 = \min(\sigma_1^2, \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2)$, $r = \frac{2\sigma_M}{\sigma_m}$. Now we state the following assumptions.

Assumption 2. *For some positive constants c_1 and d_1-d_3,*

$$C_{\min 1} \geq d_1 \left(\frac{\log p}{n} + c_1 \right) \sigma_2^2, \quad C_{\min 2} \geq \max \left(d_2 (\frac{\log p}{n} + \log r) \sigma_M^2, d_3 \frac{\log p}{n^{1/3}} \sigma_M^2 \right),$$

where $C_{\min 1} = \min_{A_1 \in \mathbb{A}} \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0||^2}{n|A \backslash A_1|}$ and $C_{\min 2} = \min_{B_1 \in \mathbb{B}} \frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B \boldsymbol{\alpha}_B^0||^2}{n|B \backslash B_1|}$.

Assumption 3. *Assume that*

$$0 < \tau_1 \leq \sigma_2 \sqrt{\frac{6}{(n+2) \cdot p \cdot c_{\max}(\mathbf{Z}^T \mathbf{Z})}},$$

and

$$0 < \tau_2 \leq \sigma_M \sqrt{\frac{6}{(n+2) \cdot p \cdot c_{\max}(\mathbf{Z}^T\mathbf{Z})}},$$

where $c_{\max}(\cdot)$ *denotes the largest eigenvalue of a matrix.*

Assumption 4. *For some positive constant* d_4*,*

$$C_{\min 3} \geq d_4 \frac{(p_0+2)\log p}{n^{1/3}} \sigma_2^2,$$

where $C_{\min 3} = \min_{B_1 \in \mathbb{G}} \frac{||\mathbf{M}_{\mathbf{Z}_{B_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0||}{n|A \backslash B_1|}$.

Assumption 2 says that $\mathbf{Z}_A\boldsymbol{\gamma}_A^0$, the imposed effect of $\mathbf{Z}_A$ on exposure $\boldsymbol{D}$, is separated from the column space spanned by $\mathbf{Z}_{A_1}$ for any A_1 that is not identical to A and contains no more than $|A|$ IVs; and $\mathbf{Z}_B\boldsymbol{\alpha}_B^0$, the imposed direct effect of $\mathbf{Z}_B$ on outcome $\boldsymbol{Y}$, is separated from the column space spanned by $\mathbf{Z}_A\boldsymbol{\gamma}_A^0$ and $\mathbf{Z}_{B_1}$ for any B_1 that is not identical to B and contains no more than $|B| = p_0$ IVs. Assumption 3 says that τ_1 and τ_2 are sufficiently small to have an adequate TLC approximation to the L_0-constraint. When $A \nsubseteq B_1$, Assumption 4 ensures that $\mathbf{Z}_A\boldsymbol{\gamma}_A^0$ is separated from the column space of $\mathbf{Z}_{B_1}$. Under Assumptions 1 to 4, Theorem 3 shows $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$ from (2.4) has the oracle property.

Theorem 3. *Under Assumptions 1 to 4, if* $K_1 = |A|$ *and* $K_2 = p_0$*, then* $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$*, either as* $n \to \infty$ *with a fixed* p*, or as* $n, p \to \infty$*.*

Theorem 3 states that, for fixed p and $n \to \infty$, or $p \to \infty$ in the order of $O\left(\exp(C \cdot n^{1/3})\right)$ for some constant C as $n \to \infty$, the constrained maximum likelihood estimator $(\hat{\beta}, \hat{\alpha})$ asymptotically has the same performance as the oracle estimator $(\hat{\beta}^{or}, \hat{\alpha}^{or})$. Together with the asymptotic variance of $\hat{\beta}^{or}$ from Theorem 2, we could perform the Wald test with the constrained estimator $(\hat{\beta}, \hat{\alpha})$.

Besides the Wald test, we can also perform the likelihood ratio test as in [43]. In the second stage, we fit another constrained linear model of $\boldsymbol{Y}$ on $\mathbf{Z}$ only under the

null hypothesis of $\beta^0 = 0$, with the same K_2 as in (2.4),

$$\hat{\boldsymbol{\alpha}}^{(0)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} ||\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\alpha}||^2 \text{ subject to } \frac{1}{\tau_2}\sum_{j=1}^{p} \min(|\alpha_j|, \tau_2) \leq K_2. \tag{2.5}$$

Then, given $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$ in (2.4), similar to [43], after profiling out the variance parameter of the error term, we define the constrained maximum likelihood ratio (CMLR) as

$$\Lambda_n = n(\log ||\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(0)}||^2 - \log ||\boldsymbol{Y} - \hat{\beta} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}||^2). \tag{2.6}$$

Corollary 1. *Assume that Assumptions 1 to 4 are met. If $K_1 = |A|$ and $K_2 = p_0$, then under the null hypothesis of $\beta^0 = 0$, Λ_n converges in distribution to χ^2_1, a chi-squared distribution with degrees of freedom 1, either as $n \to \infty$ with a fixed p, or as both $n, p \to \infty$.*

Based on Corollary 1, we can perform the likelihood ratio test. This test gives almost the same results as those from the Wald test in our simulations and real data analyses. So we will focus on the latter.

2.2.3 Tuning parameter selection with BIC

Given individual level data, we could choose K_1 and K_2 with cross-validation, but we cannot do so with summary level data. Instead, we propose using Bayesian information Criterion (BIC) [26] to select K_1 and K_2 in the two stages. For our target applications in GWAS, the sample size n is at least a few thousands, much larger than p, the number of IVs (SNPs) used, so here we consider BIC with low-dimensional data, i.e. with a fixed p. For an extension to high-dimensional data, a modified BIC such as in [7, 41] could be used but will not be pursued here. Denote the estimate from (2.3) with constraint parameter K_1 to be $\hat{\boldsymbol{\gamma}}_{K_1}$, that from (2.4) with constraint parameter

K_2 to be $(\hat{\beta}_{K_2}, \hat{\boldsymbol{\alpha}}_{K_2})$. Ignoring constant terms, log-likelihood for the first stage is

$$l_1(\hat{\boldsymbol{\gamma}}_{K_1}) = -\frac{1}{2}\left(n \cdot \log(\sigma_2^2) + \frac{||\boldsymbol{D} - \mathbf{Z}\hat{\boldsymbol{\gamma}}_{K_1}||^2}{\sigma_2^2}\right).$$

As σ_2^2 is unknown, we plug in its maximum likelihood estimate $\hat{\sigma}_2^2 = ||\boldsymbol{D} - \mathbf{Z}\hat{\boldsymbol{\gamma}}_{K_1}||^2/n$ and have the BIC for the first stage

$$\mathrm{BIC}_1(K_1) = n \cdot \log \frac{||\boldsymbol{D} - \mathbf{Z}\hat{\boldsymbol{\gamma}}_{K_1}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\gamma}}_{K_1}||_0. \tag{2.7}$$

Similarly, BIC for the second stage is

$$\mathrm{BIC}_2(K_2) = n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{K_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{K_2}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\alpha}}_{K_2}||_0. \tag{2.8}$$

We choose $\hat{K}_1 = \mathrm{argmin}_{K_1 \in \mathcal{K}_1} \mathrm{BIC}_1(K_1)$ and $\hat{K}_2 = \mathrm{argmin}_{K_2 \in \mathcal{K}_2} \mathrm{BIC}_2(K_2)$, where $\mathcal{K}_1$ and $\mathcal{K}_2$ are sets of candidate K_1's and K_2's, respectively. Theorem 4 shows that BIC is consistent for tuning parameter selection in both stages with a fixed p.

Theorem 4. *Assume that Assumptions 1 to 4 are met. Then, when p is fixed, if $|A| \in \mathcal{K}_1$ and $p_0 \in \mathcal{K}_2$, we have $P\left(\hat{K}_1 = |A|, \hat{K}_2 = p_0\right) \to 1$ as $n \to \infty$.*

2.3 Two-Sample Case

2.3.1 Oracle Estimator

Now we consider the case with the exposure and outcome data coming from two independent samples, which has largely facilitated the wide application of TWAS and MR with the availability of large-scale GWAS summary data on various traits. Suppose we only observe $\boldsymbol{Y}$ and $\mathbf{Z}$ in a sample of size n from model (2.1) and obtain an estimate $\hat{\boldsymbol{\gamma}}_A$ from another independent sample of size n_2. Now the oracle estimator

is defined as

$$\begin{aligned} \hat{\boldsymbol{D}} &= \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A, \\ (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}_B^{or}) &= \underset{\beta, \boldsymbol{\alpha}_B}{\operatorname{argmin}} ||\boldsymbol{Y} - \beta \cdot \hat{\boldsymbol{D}} - \mathbf{Z}_B \boldsymbol{\alpha}_B||^2. \end{aligned} \tag{2.9}$$

Assumption 5. *Assume that as $n_2 \to \infty$, we have $n/n_2 \to w$ for some positive and finite constant w, and $\hat{\boldsymbol{\gamma}}_A \sim N(\boldsymbol{\gamma}_A^0, \sigma_2^2 \boldsymbol{\Theta})$ with $n_2 \boldsymbol{\Theta} \to \boldsymbol{\Theta}_0$ as $n_2 \to \infty$.*

Assumption 5 states that the sample sizes of the two independent samples, n and n_2, should be in the same order, and $\hat{\boldsymbol{\gamma}}_A$ should be a consistent and asymptotically normal estimator of the true $\boldsymbol{\gamma}_A^0$. For example, we can apply the maximum likelihood or least-squares estimator to an independent sample of size n_2,

$$\boldsymbol{D}_2 = \mathbf{Z}_2 \boldsymbol{\gamma}^0 + \boldsymbol{\xi}_2. \tag{2.10}$$

Given a known A, we obtain from this sample that

$$\hat{\boldsymbol{\gamma}}_A = (\mathbf{Z}_{2,A}^T \mathbf{Z}_{2,A})^{-1} \mathbf{Z}_{2,A}^T \boldsymbol{D}_2 = \boldsymbol{\gamma}_A^0 + \boldsymbol{e}, \qquad \boldsymbol{e} \sim N\left(0, \sigma_2^2 (\mathbf{Z}_{2,A}^T \mathbf{Z}_{2,A})^{-1}\right). \tag{2.11}$$

Here $\boldsymbol{\Theta} = (\mathbf{Z}_{2,A}^T \mathbf{Z}_{2,A})^{-1}$; as usual, we assume $n_2 \boldsymbol{\Theta} = (\mathbf{Z}_{2,A}^T \mathbf{Z}_{2,A}/n_2)^{-1} \to \boldsymbol{\Theta}_0$ as $n_2 \to \infty$.

Let $\boldsymbol{\Psi}_2 = \begin{pmatrix} (\boldsymbol{\gamma}_A^0)^T \mathbf{U}_A \\ \mathbf{U}_{AB}^T \end{pmatrix} \boldsymbol{\Theta}_0 \begin{pmatrix} \mathbf{U}_A \boldsymbol{\gamma}_A^0 & \mathbf{U}_{AB} \end{pmatrix}$. Theorem 5 gives some statistical properties of the two-sample oracle estimator.

Theorem 5. *With Assumptions 1 and 5 satisfied, the probability of the oracle estimator $\hat{\beta}^{or}$ defined in (2.9) being unique converges to 1 as $n, n_2 \to \infty$, and $\hat{\beta}^{or}$ is a consistent estimator of true causal effect β^0 with $\hat{\beta}^{or} \xrightarrow{p} \beta^0$ as $n, n_2 \to \infty$. Furthermore, we have $\sqrt{n}(\hat{\beta}^{or} - \beta^0) \xrightarrow{d} N(0, v)$, with variance $v = (\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2 \sigma_2^2) \cdot (\boldsymbol{\Sigma}^{-1})_{11} + w(\beta^0)^2 \sigma_2^2 (\boldsymbol{\Sigma}^{-1} \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}^{-1})_{11}$. Under the null hypothesis H_0: $\beta^0 = 0$, we have $v = \sigma_1^2 \cdot (\boldsymbol{\Sigma}^{-1})_{11}$.*

A consistent estimator of the asymptotic variance v is

$$\hat{v}^{or} = \hat{v}_1^{or} \cdot (\hat{\boldsymbol{\Sigma}}^{or})_{11}^{-1} + \frac{n}{n_2}(\hat{\beta}^{or})^2\hat{\sigma}_2^2 \left((\hat{\boldsymbol{\Sigma}}^{or})^{-1}\hat{\boldsymbol{\Psi}}_2^{or}(\hat{\boldsymbol{\Sigma}}^{or})^{-1}\right)_{11} \xrightarrow{p} v,$$

where $\hat{v}_1^{or}$ and $\hat{\sigma}_2^2$ are consistent estimates of $(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)$ and σ_2^2, $\hat{\boldsymbol{\Sigma}}^{or}$ and $\hat{\boldsymbol{\Psi}}_2^{or}$ are consistent estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}_2$, and n and n_2 are the sample sizes for the two stages, respectively. Details are given in the Appendix.

2.3.2 New Method: Two-stage Constrained Maximum Likelihood

As for the two-sample oracle estimator, with $\hat{\boldsymbol{\gamma}}_A$ calculated from one sample, we obtain the predicted exposure as $\hat{\boldsymbol{D}} = \mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A$ for the other independent sample. Then we solve in the second stage a constrained minimization as in (2.4) to obtain the 2ScML estimator. Under Assumption 1, $\mathbf{Z}_A^T\mathbf{Z}_A/n \to \mathbf{U}_A$, so we can get a finite upper-bound of the eigenvalues of $\mathbf{Z}_A^T\mathbf{Z}_A/n$ as u_1. Similarly, from Assumption 5, we obtain a finite upper-bound of the eigenvalues of $n_2\boldsymbol{\Theta}$ as u_2, and a finite upper-bound of n/n_2 as u_3. For two-sample 2ScML, we substitute Assumption 2 by Assumption 6.

Assumption 6. *Denote* $\sigma_M^2 = \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2 + u_1u_2u_3(\beta^0)^2\sigma_2^2$, $\sigma_m^2 = \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2$, $r = \frac{2\sigma_M}{\sigma_m}$. *For some positive constants* d_2 *and* d_3,

$$C_{\min 2} \geq \max\left(d_2(\frac{\log p}{n} + \log r)\sigma_M^2, d_3\frac{\log p}{n^{1/3}}\sigma_M^2\right),$$

where $C_{\min 2} = \min_{B_1 \in \mathbb{B}} \frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n|B\backslash B_1|}$.

Theorem 6 shows the 2ScML estimator has the oracle property.

Theorem 6. *Under Assumptions 1, 3, 4, 5 and 6, if* $K_2 = p_0$, *then* $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$, *either as* $n \to \infty$ *with a fixed* p, *or as both* $n, p \to \infty$.

Similar to the one-sample case, we can also perform the likelihood ratio test for the two-sample case, with $\hat{\boldsymbol{\alpha}}^{(0)}$ and CMLR defined in (2.5) and (2.6) respectively.

Corollary 2. *Assume that Assumptions 1, 3, 4, 5 and 6 are met. If $K_2 = p_0$, then under the null hypothesis of $\beta^0 = 0$, Λ_n converges in distribution to χ^2_1, a chi-squared distribution with degrees of freedom 1, either as $n \to \infty$ with a fixed p, or as both $n, p \to \infty$.*

2.3.3 Tuning parameter selection with BIC

For the two-sample case, as we assume a consistent estimator $\hat{\boldsymbol{\gamma}}_A$ in the first stage, we consider using BIC to select K_2 in the second stage. Denote the estimate from (2.4) with K_2 as $(\hat{\beta}_{K_2}, \hat{\boldsymbol{\alpha}}_{K_2})$, then BIC has a form similar to (2.8). For a candidate set $\mathcal{K}_2$ of K_2's, we choose $\hat{K}_2 = \text{argmin}_{K_2 \in \mathcal{K}_2} \text{BIC}_2(K_2)$. Theorem 7 shows the selection consistency of BIC in the second stage with a fixed p.

Theorem 7. *Assume that Assumptions 1, 3, 4, 5 and 6 are met. Then when p is fixed, if $p_0 \in \mathcal{K}_2$, we have $P\left(\hat{K}_2 = p_0\right) \to 1$ as $n \to \infty$.*

2.4 Computation

To solve the nonconvex constrained minimization (2.3), we use a difference convex (DC) method to approximate the nonconvex constraint with a sequence of convex constraints iteratively. First, we decompose the constraint function into a difference of two convex functions:

$$\frac{1}{\tau_1} \sum_{j=1}^{p} \min(|\gamma_j|, \tau_1) = \frac{1}{\tau_1} \sum_{j=1}^{p} |\gamma_j| - \max(|\gamma_j| - \tau_1, 0). \tag{2.12}$$

Given an estimate $\gamma_j^{(m)}$ at the m^{th} iteration, the subgradient of $\max(|\gamma_j| - \tau_1, 0)$ at $|\gamma_j^{(m)}|$ is $I(|\gamma_j^{(m)}| > \tau_1)$. Then,

$$\max(|\gamma_j| - \tau_1, 0) \geq \max(|\gamma_j^{(m)}| - \tau_1, 0) + (|\gamma_j| - |\gamma_j^{(m)}|) \cdot I(|\gamma_j^{(m)}| > \tau_1). \tag{2.13}$$

Combining equations (2.12) and (2.13) yields that

$$\begin{aligned}\frac{1}{\tau_1}\sum_{j=1}^{p}\min(|\gamma_j|, \tau_1) &\leq \frac{1}{\tau_1}\sum_{j=1}^{p}|\gamma_j| - \max(|\gamma_j^{(m)}| - \tau_1, 0) - (|\gamma_j| - |\gamma_j^{(m)}|) \cdot I(|\gamma_j^{(m)}| > \tau_1) \\ &= \frac{1}{\tau_1}\sum_{j=1}^{p}|\gamma_j| \cdot I(\gamma_j^{(m)} \leq \tau_1) + \tau_1 \cdot I(\gamma_j^{(m)} > \tau_1).\end{aligned} \tag{2.14}$$

On this ground, at the m^{th} iteration, we relax the nonconvex minimization (2.3) to the following convex constrained minimization

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} ||\boldsymbol{D} - \mathbf{Z}\boldsymbol{\gamma}||^2 \text{ subject to } \frac{1}{\tau_1}\sum_{j=1}^{p}|\gamma_j| \cdot I(\gamma_j^{(m)} \leq \tau_1) \leq K_1 - \sum_{j=1}^{p} I(\gamma_j^{(m)} > \tau_1). \tag{2.15}$$

Solving (2.15) is equivalent to solving a constrained lasso problem, which can be solved by an algorithm in [21]. Similarly, we iteratively relax nonconvex minimization (2.4) to

$$\begin{aligned}&(\hat{\beta}, \hat{\boldsymbol{\alpha}}) = \underset{\beta, \boldsymbol{\alpha}}{\operatorname{argmin}} ||\boldsymbol{Y} - \beta \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\boldsymbol{\alpha}||^2, \text{ subject to} \\ &\frac{1}{\tau_2}\sum_{j=1}^{p}|\alpha_j| \cdot I(\alpha_j^{(m)} \leq \tau_2) \leq K_2 - \sum_{j=1}^{p} I(\alpha_j^{(m)} > \tau_2).\end{aligned} \tag{2.16}$$

Again, the algorithm of [21] is applied to solve the constrained lasso problem (2.16). This process continues until a termination criterion is met.

2.5 Extension to GWAS Summary Data

We extend the proposed 2ScML to a situation with only GWAS summary statistics and a reference panel. For a certain trait or outcome Y, a GWAS is performed to assess possible associations between the SNPs and Y. By performing marginal linear regression of Y on each SNP Z separately, we estimate the marginal effect size of Z on Y as $\hat{\beta}_{YZ}$ along with its standard error $\text{se}(\hat{\beta}_{YZ})$. Due to the logistic and privacy issues, individual-level genotypes (i.e. Z's) and phenotypes (i.e. Y) are typically not publicly available but only summary data in the form of $\hat{\beta}_{YZ}$'s and $\text{se}(\hat{\beta}_{YZ})$'s are available for all SNPs/Z's. From a reference panel, consisting of individual-level genotype data of a group of individuals, such as from the 1000 Genomes Project [1], we estimate the correlation structure among the SNPs. This generalization allows our method to apply to some published large-scale GWAS summary data with a wide range of traits to boost the power of statistical analysis.

For the constrained lasso problem in (2.15), it can be written as:

$$||\boldsymbol{D} - \mathbf{Z}\boldsymbol{\gamma}||^2 = \boldsymbol{D}^T\boldsymbol{D} - 2\boldsymbol{D}^T\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}.$$

Recall that with individual level data, $\boldsymbol{D} \in \mathbb{R}^n$ is a vector of the exposure values for n individuals and $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is the matrix of p IVs for n individuals. Without loss of generality, we assume $\boldsymbol{D}$ and column vectors of $\mathbf{Z}$ are all standardized to have sample mean 0 and sample variance 1, so $\boldsymbol{D}^T\boldsymbol{D} = n$ and $\boldsymbol{D}^T\mathbf{Z}/n \in \mathbb{R}^p$ is a vector of correlations between exposure and p IVs, $\mathbf{Z}^T\mathbf{Z}/n \in \mathbb{R}^{p\times p}$ is the correlation matrix of p IVs. From the summary statistics of the exposure, we extract the vector of correlations between exposure D and p IVs, denoted by $\boldsymbol{r}_{DZ} \in \mathbb{R}^p$. From the reference panel consisting of n_0 individuals, we extract the values of p IVs denoted by $\mathbf{Z}_0 \in \mathbb{R}^{n_0\times p}$ and estimate their correlation matrix as $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{p\times p}$. Now in (2.15), we

replace $\boldsymbol{D}^T\mathbf{Z}/n$ and $\mathbf{Z}^T\mathbf{Z}/n$ with their estimates $\boldsymbol{r}_{DZ}$ and $\boldsymbol{\Sigma}_0$ respectively. Then

$$\begin{aligned}&\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}(1 - 2\boldsymbol{r}_{DZ}^T\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\boldsymbol{\Sigma}_0\boldsymbol{\gamma}), \text{ subject to} \\ &\frac{1}{\tau_1}\sum_{j=1}^{p}|\gamma_j| \cdot I(\gamma_j^{(m)} \leq \tau_1) \leq K_1 - \sum_{j=1}^{p} I(\gamma_j^{(m)} > \tau_1).\end{aligned} \tag{2.17}$$

When $\boldsymbol{\Sigma}_0$ is positive definite, which could be achieved by pruning out highly correlated SNPs in the reference panel, $\left(1 - 2\boldsymbol{r}_{DZ}^T\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\boldsymbol{\Sigma}_0\boldsymbol{\gamma}\right)$ is bounded below and thus has a finite minimum. Denote $\boldsymbol{\Sigma}_0^{1/2}$ as the square-root of $\boldsymbol{\Sigma}_0$ such that $\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_0^{1/2} = \boldsymbol{\Sigma}_0$. Then $\boldsymbol{\Sigma}_0^{1/2}$ is also positive definite with inverse $\boldsymbol{\Sigma}_0^{-1/2}$. Now, solving (2.17) amounts to solving

$$\begin{aligned}&\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\,||\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{r}_{DZ} - \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\gamma}||^2, \text{ subject to} \\ &\frac{1}{\tau_1}\sum_{j=1}^{p}|\gamma_j| \cdot I(\gamma_j^{(m)} \leq \tau_1) \leq K_1 - \sum_{j=1}^{p} I(\gamma_j^{(m)} > \tau_1),\end{aligned} \tag{2.18}$$

which is a constrained lasso problem. Given $\hat{\boldsymbol{\gamma}}$, we obtain $\hat{\boldsymbol{D}}$ as a function of the SNPs. Similarly, we obtain the correlations between Y and the SNPs, and between Y and $\hat{D}$ from the summary statistics for outcome Y, and solve (2.16) as before in the second stage.

In the first stage, given K_1, we obtain an estimate $\hat{\boldsymbol{\gamma}}_{K_1}$. Paralleling with (2.7), the BIC is

$$\text{BIC}_1(\hat{\boldsymbol{\gamma}}_{K_1}) = n \cdot \log\left(1 - 2\boldsymbol{r}_{DZ}^T\hat{\boldsymbol{\gamma}}_{K_1} + \hat{\boldsymbol{\gamma}}_{K_1}^T\boldsymbol{\Sigma}_0\hat{\boldsymbol{\gamma}}_{K_1}\right) + \log(n) \cdot ||\hat{\boldsymbol{\gamma}}_{K_1}||_0. \tag{2.19}$$

Similarly, we select K_2 using the BIC as (2.8) in the second stage.

When we draw inference about the causal effect of one trait on another, typically GWAS summary data for both the exposure and outcome are available. In TWAS applications, for each gene, the exposure D is its expression level, the IVs are some eSNPs for this gene, and Y is some trait such as LDL. Often a GWAS summary

dataset for gene expression (i.e. so-called eQTL data) is not available, so we cannot perform the first stage analysis with 2SLS or 2ScML. Instead, some pre-calculated $\hat{\boldsymbol{\gamma}}$'s, e.g. estimated from some publicly unavailable eQTL data by penalized regression like Lasso or Elastic Net, are publicly available as from the TWAS Fusion website [15]. Our proposed 2ScML method here was motivated and is thus applicable with these pre-calculated $\hat{\boldsymbol{\gamma}}$'s, from which we obtain $\hat{\boldsymbol{D}}$ in the second stage of 2ScML (or 2SLS) with some GWAS summary data for Y as to be shown in our TWAS applications.

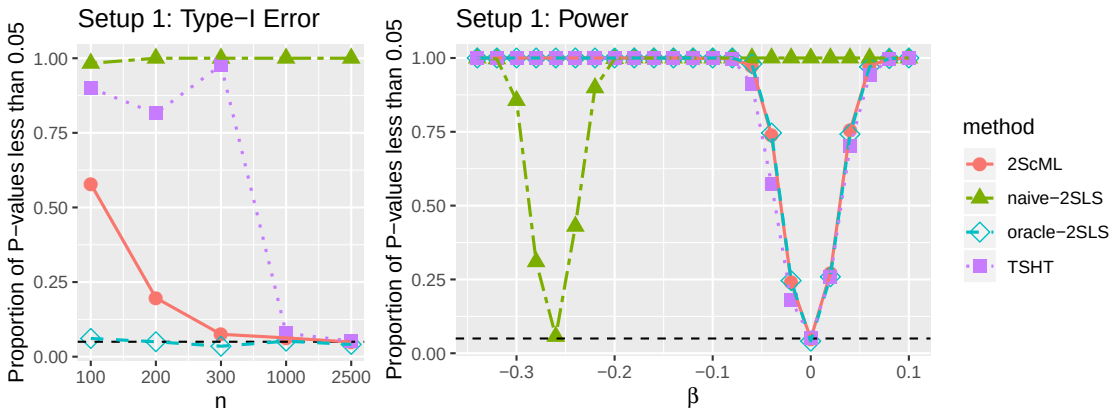
Chapter 3

Simulations

3.1 Simulation 1: One-Sample Case with IV Assumptions (A) and (C) Violated

We compare 2ScML with TSHT, naive-2SLS/TWAS, and oracle-2SLS through simulations, and the simulation setups closely follow those of TSHT [14]. Since TSHT, as a state-of-the-art method, applied only to the one-sample case, we consider the IV Assumptions (A) and (C) being violated in such a case. First we compare the Type-I error rates of these methods. In Setup 1, we set the number of IVs $p = 100$, and the sample size n varying at $100, 200, 300, 1000$ and 2500. Instruments $\boldsymbol{Z}$'s followed a multivariate normal distribution with mean 0 and an AR(0.5) covariance matrix. The error terms (ϵ, ξ) were generated from a bivariate normal distribution with mean 0, variances 1.5 and covariance 0.75. For $2 \leq i \leq 8$, $\gamma_i^0 = 0.5$; otherwise $\gamma_i^0 = 0$; i.e. the 2^{nd} to 8^{th} IVs were relevant with an equal effect size 0.5. For $i = 7, 8$, $\alpha_i^0 = 0.5$; otherwise $\alpha_i^0 = 0$; i.e. the 7^{th} and 8^{th} instruments were invalid IVs with some direct effects on the outcome. When $\beta^0 = 0$, there was no causal effect from the exposure to outcome, i.e. the null case. Setup 2 was similar to Setup 1 except: for $i = 1, 7, 8, 9$, $\alpha_i^0 = 0.5$; otherwise $\alpha_i^0 = 0$; i.e. the relevant 7^{th} and 8^{th} instruments were invalid, and the irrelevant 1^{st} and 9^{th} instruments were also invalid. Since the results for Setup 2

Figure 3.1: Empirical Type-I error rates (left panel) and power (right panel) for Simulation 1 (Setup 1) at the nominal level 0.05.



were similar to those of Setup 1, we relegate the details to the Appendix.

In each simulation we generated n samples from the model in (2.1), then we applied the four methods to the simulated data to test $H_0 : \beta^0 = 0$ versus $H_1 : \beta^0 \neq 0$. For 2ScML, in the first stage, we used BIC to choose the best $K_1 \in \{5, 6, 7, 8, 9, 10\}$; in the second stage, we used BIC to choose the best $K_2 \in \{0, 1, 2, 3, 4, 5\}$; and we set $\tau_1 = \tau_2 = 1 \times 10^{-5}$. For the naive-2SLS, in the first stage we used the 2^{nd} to 8^{th} IVs to get $\hat{D}$, then in the second stage we fitted a linear regression model of Y on $\hat{D}$. The oracle-2SLS had the same model as that of the naive-2SLS in the first stage, but in the second stage we fitted a linear regression model of Y on $\hat{D}$ and the 7^{th} and 8^{th} IVs for Setup 1, or $\hat{D}$ and the 1^{st}, 7^{th}, 8^{th} and 9^{th} IVs for Setup 2; in other words, we included the 2 and 4 invalid IVs with direct effects in the stage 2 models for Setup 1 and Setup 2 respectively.

For each setup, we repeated the simulation 1000 times and set the nominal significance level at 0.05 for each n; Figure 3.1 summarizes the simulation result. We can see that the oracle-2SLS could always have a Type-I error rate around the nominal level 0.05, while naive-2SLS had a Type-I error rate dramatically inflated around 1. When the sample size was small, TSHT and 2ScML both had large Type-I error rates;

but as the sample size increased from 100 to 300, the Type-I error rate of 2ScML decreased fast, while that of TSHT remained large. When the sample size was large enough, both TSHT and 2ScML could control the Type-I error rate at 0.05. It is noted that, due to $\alpha_7^0 = \alpha_8^0 = \gamma_7^0 = \gamma_8^0 = 0.5$ for the two invalid IVs, it required larger sample sizes for the two consistent estimation methods to distinguish the direct and indirect effects of the IVs and thus maintain a correct Type-I error rate.

As to be seen from the left panel of Figure 3.1, when the sample size was 2500, 2ScML, TSHT, and oracle-2SLS could control their Type-I error rates. So we compare their power at the sample size 2500: we changed β^0 from -0.5 to 0.1 with a step size of 0.02 and applied all four methods with 1000 independent replicates to calculate their empirical power. Figure 3.1 (right panel) shows the results. Again, when $\beta^0 = 0$, i.e. with no causal effect, 2ScML, TSHT, and oracle-2SLS could control Type-I error at 0.05, while naive-2SLS had its Type-I error inflated to 1. When $|\beta^0|$ was large, 2ScML, TSHT, and oracle-2SLS all had power 1. When $|\beta^0|$ was small, the power of 2ScML and oracle-2SLS was very close and typically higher, and sometimes much higher, than that of TSHT. For example, for $\beta^0 = -0.04$, the empirical power of the oracle and 2ScML was close at 0.746 and 0.738 respectively, much higher than that of TSHT at 0.572. This supports the theory that 2ScML has the oracle property while TSHT does not. We can see that for the range $-0.3 < \beta^0 < -0.22$, naive-2SLS had power smaller than 1, while the other three methods had power 1. This was not surprising: the direct effects $\boldsymbol{\gamma}^0$ and $\boldsymbol{\alpha}^0$ were positive, and the total effect of the IVs on Y was $\beta^0 \cdot \boldsymbol{\gamma}^0 + \boldsymbol{\alpha}^0$, so negative β^0's would diminish the total effect toward 0 and thus decrease the power of naive-2SLS.

3.2 Simulation 2: Two-Sample Case with IV Assumption (B) Violated

We compare 2ScML with perhaps the most popular method to deal with invalid IVs in MR, MR-Egger regression [4], with simulated data. As required by MR, we consider the two-sample case. Here we used p IVs Z_1 to Z_p, all of which were relevant to exposure D (i.e. IV Assumption A was satisfied), but some of Z's had direct effects on outcome Y, and some of Z's might have direct effects on unobserved confounder U. The data-generating model was:

$$\begin{aligned} U_i &= \boldsymbol{Z}_i\boldsymbol{\phi}^0 + \epsilon_{U,i}, \\ D_i &= \boldsymbol{Z}_i\boldsymbol{\gamma}^0 + U_i + \epsilon_{D,i}, \\ Y_i &= \beta^0 \cdot D_i + \boldsymbol{Z}_i\boldsymbol{\alpha}^0 + U_i + \epsilon_{Y,i}. \end{aligned} \tag{3.1}$$

Here $\epsilon_U, \epsilon_D, \epsilon_Y$ were independent random errors. $\boldsymbol{\phi}^0$ were direct effects from Z's to U, $\boldsymbol{\gamma}^0$ were direct effects from Z's to D, $\boldsymbol{\alpha}^0$ were direct effects from Z's to Y; β^0 was the causal effect from D to Y, the parameter of interest. MR-Egger regression uses summary statistics of Z's on D, and Z's on Y, from two independent datasets. It requires Z's are independent IVs, and all Z's are relevant for D, i.e. $\boldsymbol{\gamma}^0$ are all non-zero. $\boldsymbol{\alpha}^0$ could be non-zero, i.e. IV assumption (C) could be violated, but $\boldsymbol{\alpha}^0$ should be independent of $\boldsymbol{\gamma}^0$; this is the so-called InSIDE assumption required by MR-Egger regression. Accordingly, it does not allow non-zero $\boldsymbol{\phi}^0$, i.e. the IV assumption (B) should be satisfied. In contrast, as to be confirmed, our new method 2ScML allows IV Assumption (B) to be violated, i.e. non-zero $\boldsymbol{\phi}^0$.

We generated simulated data similar to the Case 3 with directional pleiotropy and violated InSIDE assumption in [29]. We set the sample size $n = 20000$, generated $p = 30$ independent SNPs as Z's with minor allele frequency (MAF) 0.3. We generated

$\epsilon_U, \epsilon_D, \epsilon_Y$ from a standard normal distribution with mean 0 and standard deviation 1. $\boldsymbol{\gamma}^0$'s were generated from truncated normal distribution with mean 0 and standard deviation 0.2, and truncate at 0.1, i.e. all $\boldsymbol{\gamma}^0$'s having absolute values larger than 0.1; this was used to satisfy IV assumption (A). We had 30% invalid IVs, and we chose them as the first 9 Z's, and generated $\alpha_1^0, \cdots, \alpha_9^0$ from a normal distribution with mean 0.5 and standard deviation 0.075. $\phi_1^0, \cdots, \phi_9^0$ were drawn from normal distribution with mean 0 and standard deviation σ_ϕ. When $\sigma_\phi = 0$, $\phi_1^0, \cdots, \phi_9^0$ were all 0's, implying IV Assumption (B) was satisfied; when $\sigma_\phi \neq 0$, $\phi_1^0, \cdots, \phi_9^0$ were non-zero, leading to IV Assumption (B) violated. We tried $\sigma_\phi = 0$ or 0.1, and $\beta^0 \in \{-0.1, -0.05, -0.02, -0.01, 0, 0.01, 0.02, 0.05, 0.1\}$. In each simulation, we generated two independent samples.

For MR-Egger regression, we calculated the summary statistics for D in the first sample and for Y in the second sample, and used function `mr_egger_regression()` from R package **TwoSampleMR**. For 2ScML, we used the first sample to calculate $\hat{\boldsymbol{\gamma}}$ with linear regression and all 30 Z's, then used $\hat{\boldsymbol{\gamma}}$ to get $\hat{D}$ in the second sample, and performed the second stage with $\tau_2 = 1 \times 10^{-5}$ and chose the best K_2 from $7, 8, 9, 10, 11$ with BIC, as the true number of invalid Z's was 9. We also applied the Oracle estimator, which had the same first stage as 2ScML, and in the second stage, we included $\hat{D}$ and the first 9 Z's. For each setup, we did simulation 1000 times, and calculated empirical Type-I error rates (for $\beta^0 = 0$) and power (for $\beta^0 \neq 0$). Figure 3.2 shows the simulation results.

From Figure 3.2 we can see that, in the upper panel when $\sigma_\phi = 0$, i.e. when IV Assumption (B) and thus the InSIDE assumption was satisfied, all three methods could control the Type-I error rates well at the nominal level of 0.05 for $\beta^0 = 0$; for $\beta^0 \neq 0$, 2ScML and Oracle had similar power, much higher than MR-Egger. In the lower panel when $\sigma_\phi = 0.1$, i.e. when IV Assumption (B) and thus the InSIDE assumption was violated, MR-Egger could not control its Type-I error rate, while

Figure 3.2: Simulation 2: empirical Type I error rates (for $\beta^0 = 0$) and power (for $\beta^0 \neq 0$) when IV Assumption (B) was not (upper panel) or was (lower) violated.

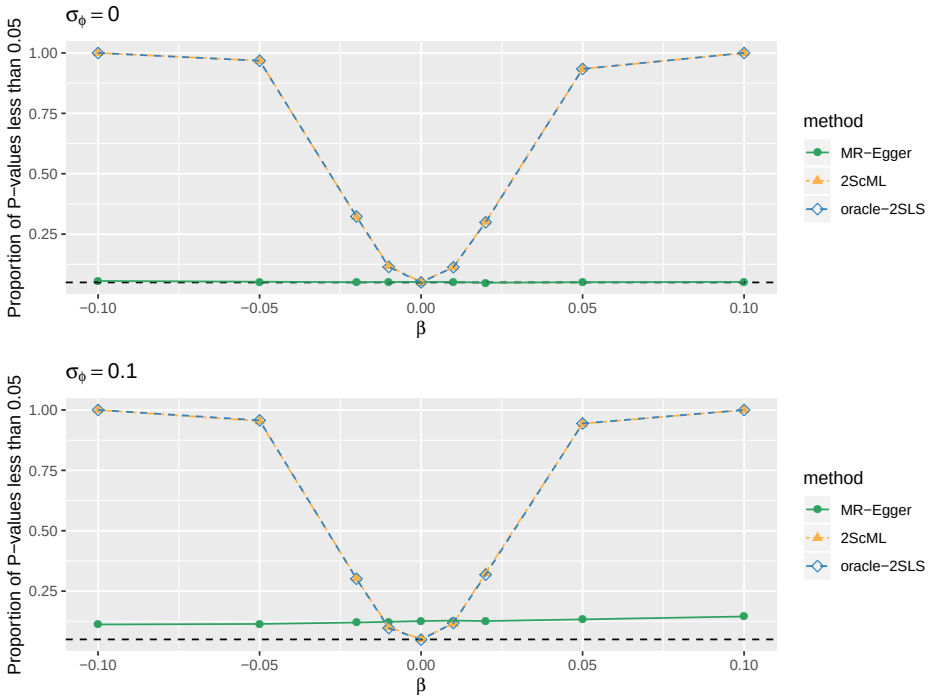


Table 3.1: The means ($\bar{\hat{\beta}}$ and $\bar{SE}(\hat{\beta})$) of the estimates $\hat{\beta}$ and their standard errors, and the standard deviations (sd) of $\hat{\beta}$ for MR-Egger, 2ScML and the oracle methods in Simulation 2 with $\sigma_\phi = 0$ (i.e. with IV Assumption (B) and thus the InSIDE assumption holding).

β^0	MR-Egger			2ScML			Oracle		
	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$
-0.10	-0.0959	0.5370	0.5220	-0.1001	0.0132	0.0131	-0.1001	0.0131	0.0131
-0.05	-0.0466	0.5366	0.5219	-0.0503	0.0134	0.0134	-0.0503	0.0133	0.0134
-0.02	-0.0170	0.5364	0.5218	-0.0204	0.0136	0.0136	-0.0204	0.0135	0.0136
-0.01	-0.0071	0.5363	0.5217	-0.0104	0.0136	0.0136	-0.0104	0.0136	0.0136
0	0.0027	0.5362	0.5217	-4e-04	0.0137	0.0137	-4e-04	0.0136	0.0137
0.01	0.0126	0.5361	0.5217	0.0096	0.0138	0.0138	0.0095	0.0137	0.0138
0.02	0.0224	0.5360	0.5216	0.0196	0.0138	0.0138	0.0195	0.0137	0.0138
0.05	0.0519	0.5357	0.5215	0.0494	0.0141	0.0141	0.0494	0.0140	0.0141
0.10	0.1011	0.5353	0.5213	0.0993	0.0145	0.0145	0.0993	0.0144	0.0145

2ScML and Oracle could. Again 2ScML and Oracle had similar and much higher power than MR-Egger.

Table 3.1 and Table 3.2 show the results in terms of estimating the causal effect size β^0. For $\sigma_\phi = 0$, i.e. when the InSIDE assumption was satisfied, MR-Egger was nearly unbiased, though the standard error (SE) of its estimate was much larger than those from the other two methods, which were quite close to each other, explaining their power properties as shown in Figure 3.2. When $\sigma_\phi = 0$, i.e. the InSIDE assumption was violated, the bias of MR-Egger was large. In contrast, 2ScML and Oracle always gave negligibly small biases. For all three methods, when $\sigma_\phi = 0$, the mean of the SE estimates of $\hat{\beta}$ was close to the corresponding standard deviation of $\hat{\beta}$, sd($\hat{\beta}$), though MR-Egger had much larger standard errors than those of 2ScML and Oracle, which were quite close to each other, explaining their empirical Type I error rate and power properties as shown in Figure 3.2. When $\sigma_\phi = 0.1$, the means of SE($\hat{\beta}$) remained close to sd($\hat{\beta}$) for 2ScML and Oracle, but were quite different for MR-Egger.

Table 3.2: The means ($\bar{\hat{\beta}}$ and $\bar{SE}(\hat{\beta})$) of the estimates $\hat{\beta}$ and their standard errors, and the standard deviations (sd) of $\hat{\beta}$ for MR-Egger, 2ScML and the oracle methods in Simulation 2 with $\sigma_\phi = 0.1$ (i.e. with IV Assumption (B) and thus the InSIDE assumption violated).

β^0	MR-Egger			2ScML			Oracle		
	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$	$\bar{\hat{\beta}}$	sd($\hat{\beta}$)	$\bar{SE}(\hat{\beta})$
-0.10	0.0662	0.5745	0.4629	-0.0990	0.0132	0.0130	-0.0990	0.0130	0.0130
-0.05	0.1163	0.5741	0.4628	-0.0491	0.0134	0.0133	-0.0491	0.0132	0.0133
-0.02	0.1463	0.5738	0.4627	-0.0192	0.0135	0.0134	-0.0192	0.0133	0.0134
-0.01	0.1562	0.5737	0.4627	-0.0093	0.0135	0.0135	-0.0093	0.0134	0.0135
0	0.1662	0.5736	0.4627	7e-04	0.0136	0.0136	7e-04	0.0134	0.0136
0.01	0.1762	0.5735	0.4626	0.0107	0.0137	0.0136	0.0107	0.0135	0.0136
0.02	0.1862	0.5734	0.4626	0.0207	0.0137	0.0137	0.0206	0.0136	0.0137
0.05	0.2161	0.5731	0.4625	0.0506	0.0139	0.0139	0.0505	0.0138	0.0139
0.10	0.2658	0.5726	0.4624	0.1004	0.0144	0.0144	0.1004	0.0142	0.0144

3.3 Other Simulations

In Appendix, we show the results from Simulation 3, demonstrating the good performance of the proposed method with a realistic set-up using real genotypic data for a gene and mimicking real TWAS data.

Chapter 4

Real Data Example

We applied 2ScML and the standard 2SLS/TWAS to identify (putative) causal genes for LDL with GWAS summary statistics. Note that TSHT does not apply to GWAS summary data, and thus cannot be compared here. For each gene, we used the TWAS Fusion pre-calculated coefficients $\hat{\boldsymbol{\gamma}}$'s for our first stage analysis [15]; the coefficients were estimated with elastic net regression based on microarray expression data of blood from the Young Finns Study (YFS) with sample size 1264 [20, 23]. From the 1000 Genomes Project [1], we took 489 unrelated individuals of European ancestry as our reference panel. The GWAS summary data of LDL were drawn from [32] with sample sizes up to 95454; we removed the SNPs with sample sizes less than 80000. We used software `ImpG` [22] to impute with the reference panel for the LDL GWAS summary statistics. As stated in [22], we used the imputation accuracy measure r^2 to quantify the imputation quality for each SNP and removed imputed SNPs with $r^2 < 0.3$.

There were 4700 genes with pre-calculated $\hat{\boldsymbol{\gamma}}$ in the TWAS Fusion database. For each gene, we first identified the set of its eSNPs with non-zero coefficients $\hat{\gamma}$. We removed 120 genes with less than half of their eSNPs (with non-zero coefficients) present in both the reference panel and GWAS summary data. We also pruned these eSNPs to make their pairwise absolute correlations no larger than 0.9, and used the

remaining p eSNPs to predict each gene's expression as $\hat{D}$ in the first stage analysis. In the second stage we included $\hat{D}$ and its eSNPs, and used BIC to select the best K_2 from $\{1, 2, \cdots, \lceil (p-1)/2 \rceil\}$.

For each of the 4580 genes, we obtained its p-values from the standard TWAS and our method, denoted as p_{TWAS} and p_{2ScML} respectively. After the Bonferroni correction, we obtained 32 significant genes with at least one of their p_{TWAS} and p_{2ScML} less than 0.05/4580; the standard TWAS and our new method identified 21 and 23 significant genes respectively, including 12 common ones. We did a literature search on each of the 32 significant genes. We excluded the study generating the LDL GWAS data we used [32]. Based on the literature support from other studies, we assigned a score to each gene: if there were other studies (1) supporting this gene being associated with LDL, we assigned the highest score of 5; (2) supporting this gene associated with a trait related to LDL, we assigned a score of 4; (3) identifying one or more SNPs mapped to or nearby this gene, which were significantly associated with LDL, we assign a score of 3; (4) identifying some SNPs mapped to or nearby this gene, which were significantly associated with other traits related to LDL, we assigned it a score of 2; (5) identifying some SNPs mapped to or nearby this gene that were significantly associated with any traits, we assigned a score of 1; (6) otherwise, we assigned the lowest score of 0. See Appendix for a list of all the 32 genes with their supporting references.

We compare the two sets of the p-values of the 32 genes obtained by the two methods in Figure 4.1. For better visualization, we show the p-values in the $-\log_{10}$ scale. The left panel gives a scatter plot for all the 32 genes, while the right panel zooms in; the vertical and horizontal green dashed lines represent the Bonferroni adjusted significance cutoff 0.05/4580. The larger a score assigned to a gene (i.e. with stronger literature support), the larger the size and the darker the color of its corresponding point in the plot. Table 4.1 shows the 8 genes with a score of 4 or 5. It is

Figure 4.1: Comparison of the p-values in the $-\log_{10}$ scale of the 32 significant genes for LDL.

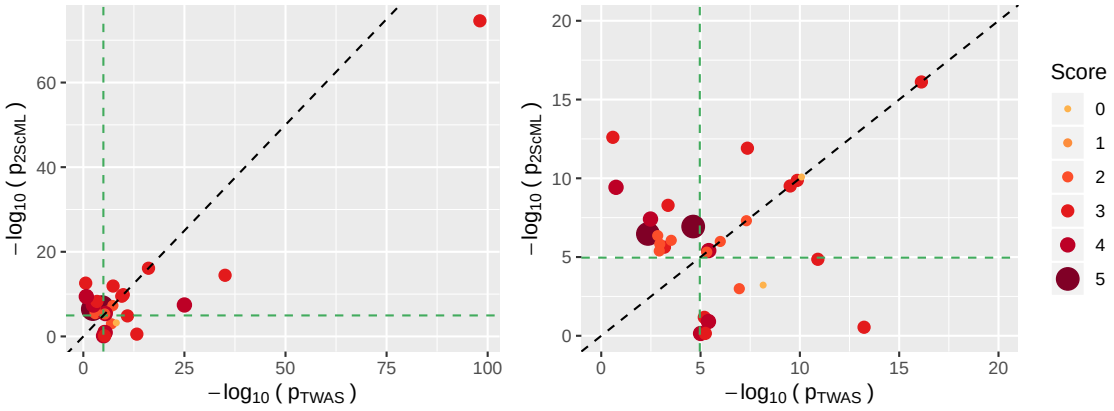


Table 4.1: The eight genes each with a score of 4 or 5 and identified by TWAS or 2ScML to be associated with LDL.

Gene	Chromosome	p	Best K_2	p_{TWAS}	p_{2ScML}	Score
LDLRAP1	1	13	1	2.35e-05	1.15e-07	5
GNAI3	1	26	8	9.93e-06	7.11e-01	4
CCDC93	2	3	1	4.46e-03	3.39e-07	5
DDAH2	6	13	0	3.87e-06	3.87e-06	4
HP	16	14	1	3.31e-03	3.84e-08	4
CARM1	19	9	2	4.02e-06	1.24e-01	4
SMARCA4	19	3	1	1.01e-25	3.62e-08	4
LPAR2	19	6	1	1.80e-01	3.73e-10	4

notable that the following four genes with strong literature support for their relevance to LDL were identified by our new method, but missed by the standard TWAS: both genes *LDLRAP1* and *CCDC93* were reported to be associated with LDL [40, 13, 24]; gene *HP* was linked to diabetic nephropathy [2] and incidence of coronary artery disease in type 1 diabetes [25]; gene *LPAR2* was reported as a potential effector gene for fatty liver [11].

Chapter 5

Conclusion and Discussion

We have proposed a Two-Stage Constrained Maximum Likelihood (2ScML) method to draw inference on causal effects in the presence of invalid instruments. Here, in addition to allowing correlated and high-dimensional IVs, we allow the violation of some or all of the three IV assumptions; our modeling assumptions are far more general than many existing methods, such as the popular MR-Egger regression. Theoretical and simulation results confirm the oracle property of 2ScML with superior performance over the standard/naive 2SLS/TWAS, Two-Stage Hard Threshold (TSHT) and MR-Egger regression. To meet the urgent need in current genetics research (while overcoming some limitations of many existing robust IV methods that do not apply to two-sample GWAS summary data), we have developed 2ScML for both one-sample and two-sample cases and GWAS summary data, making it widely applicable to identify causal genes in TWAS and infer causal relationships between pairs of complex traits. We have applied 2ScML to a real dataset to discover causal genes for LDL with GWAS summary data, leading to some encouraging results. More applications to other data and comparisons with other existing methods warrant future investigation.

An R package implementing 2ScML with some example data, code and tutorials is publicly available at `https://github.com/xue-hr/2ScML`.

References

[1] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74.

[2] Asleh, R., Levy, A. P. (2005). In vivo and in vitro studies establishing haptoglobin as a major susceptibility gene for diabetic vascular disease. *Vascular health and risk management*, 1(1), 19.

[3] Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, Kraft P. (2018). Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genetic Epidemiology*, 42(5), 418-433.

[4] Bowden J, Davey Smith G, Burgess S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*, 44(2), 512-525.

[5] Burgess S, et al. (2017). Sensitivity analysis for robust causal inference from Mendelian randomization analysis with multiple genetic variants. *Epidemiology*, 28, 30-42.

[6] Cai, M., Chen, L., Liu, J., Yang, C. (2019). Quantifying the impact of genetically regulated expression on complex traits and diseases. bioRxiv.

[7] Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3), 759-771.

[8] Davey Smith G, Ebrahim S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32, 1-22.

[9] Davey Smith G, Ebrahim S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33, 30-42.

[10] Davey Smith G, Hemani G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23, R89-R98.

[11] DiStefano, J. K., Kingsley, C., Wood, G. C., Chu, X., Argyropoulos, G., Still, C. D., ..., Gerhard, G. S. (2015). Genome-wide analysis of hepatic lipid content in extreme obesity. *Acta diabetologica*, 52(2), 373-382.

[12] Gamazon, E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091-1098.

[13] Garcia, C. K., Wilund, K., Arca, M., Zuliani, G., Fellin, R., Maioli, M., ..., Barnes, R. (2001). Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science*, 292(5520), 1394-1398.

[14] Guo, Z., Kang, H., Tony Cai, T., Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B*, 80(4), 793-815.

[15] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., ..., Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3), 245-252.

[16] Hu, Y., Li, M., Lu, Q.,Weng, H.,Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S. et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet*, 51, 568-576.

[17] Kang, H., Zhang, A., Cai, T. T., Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513), 132-144.

[18] Lin, W., Feng, R., Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509), 270-288.

[19] Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet*, 51, 675-682.

[20] Nuotio, J., Oikonen, M., Magnussen, C. G., Jokinen, E., Laitinen, T., Hutri-Kahonen, N., ..., Jula, A. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. *Scandinavian Journal of Public Health*, 42(7), 563-571.

[21] Osborne, M. R., Presnell, B., Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2), 319-337.

[22] Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., ..., Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20), 2906-2914.

[23] Raitakari, O. T., Juonala, M., Ronnemaa, T., Keltikangas-Jarvinen, L., Rasanen, L., Pietikainen, M., ..., Jula, A. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. *International Journal of Epidemiology*, 37(6), 1220-1226.

[24] Rimbert, A., Dalila, N., Wolters, J. C., Huijkman, N., Smit, M., Kloosterhuis, N., ..., Biobank-Based Integrative Omics Studies Consortium. (2020). A common variant in CCDC93 protects against myocardial infarction and cardiovascular mortality by regulating endosomal trafficking of low-density lipoprotein receptor. *European heart journal, 41*(9), 1040-1053.

[25] Sadrzadeh, S. H., Bozorgmehr, J. (2004). Haptoglobin phenotypes in health and disorders. *Pathology Patterns Reviews, 121*(suppl_1), S97-S104.

[26] Schwarz, G. (1978). Estimating the dimension of a model. Annals of statistics, 6(2), 461-464.

[27] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association.* 107, 223-232.

[28] Shen, X., Pan, W., Zhu, Y., Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5), 807-832.

[29] Slob, E. A., Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. *Genetic Epidemiology*, 44(4), 313-329.

[30] Su YR, Di C, Bien S, Huang L, Dong X, Abecasis G, et al. (2018). A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. *Am J Hum Genet*, 102(5), 904-919.

[31] Tchetgen, Eric J. Tchetgen, BaoLuo Sun, and Stefan Walter. (2017). The GENIUS approach to robust Mendelian randomization inference. arXiv preprint arXiv:1709.07779.

[32] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., ... & Johansen, C. T. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307), 707-713.

[33] Verbanck, M., Chen, C.-Y., Neale, B., Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50, 693-698.

[34] Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., âĂę Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4), 592-599.

[35] Watanabe, K., Stringer, S., Frei, O. et al. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet*, 51, 1339-1348.

[36] Windmeijer, F., Farbmacher, H., Davies, N., Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527), 1339-1350.

[37] Wu, C., Pan, W. (2020). A powerful fine-mapping method for transcriptome-wide association studies. *Hum Genet*, 139, 199-213.

[38] Xu Z, Wu C, Wei P, Pan W. (2017a). A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics*, 207, 893-902.

[39] Xu Z, Wu C, Pan W; Alzheimer's Disease Neuroimaging Initiative. (2017b). Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Neuroimage*, 159, 159-169.

[40] Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., ..., Yin, Y. (2012). Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell research, 22*(1), 107-126.

[41] Zhang, Y., Shen, X. (2010). Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, 3(5), 350-358.

[42] Zhao Q, Wang J, Hemani G, Bowden J, Small DS. (2020). Statistical inference in two-sample summary-data Mendelian randomization using a robust adjusted profile score. *Ann Statist*, 48, 1742-1769.

[43] Zhu, Y., Shen, X., Pan, W. (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529), 217-230.

[44] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5), 481âĂŞ7.

[45] Zhu, Z., Zheng, Z., Zhang, F. et al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun*, 9, 224.

[46] Zou, H., Hastie, T., Tibshirani, R. (2007). On the “degrees of freedom" of the lasso. *Annals of Statistics*, 35(5), 2173-2192.

Appendix A

Proofs

Theorem 2. *With Assumption 1 satisfied, the probability of the oracle estimator $\hat{\beta}^{or}$ defined in (2) being unique converges to 1 as $n \to \infty$, and $\hat{\beta}^{or}$ is a consistent estimator of the true causal effect β^0 with $\hat{\beta}^{or} \xrightarrow{p} \beta^0$ as $n \to \infty$. Furthermore, we have $\sqrt{n}(\hat{\beta}^{or} - \beta^0) \xrightarrow{d} N(0, v)$, with variance $v = (\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2) \cdot (\mathbf{\Sigma}^{-1})_{11} - (2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2) \cdot (\mathbf{\Sigma}^{-1}\mathbf{\Psi}\mathbf{\Sigma}^{-1})_{11}$. Under the null hypothesis H_0: $\beta^0 = 0$, we have $v = \sigma_1^2 \cdot (\mathbf{\Sigma}^{-1})_{11}$. Here $\sigma_1^2, \sigma_2^2, \sigma_{12}$ are the variances and covariance of the error terms as defined in (1).*

Proof. We can get $\hat{\boldsymbol{\gamma}}_A^{or} = (\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{D}$ and $\hat{\boldsymbol{D}} = \mathbf{P}_{\mathbf{Z}_A}\boldsymbol{D} = \mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}$. Note that

$$\begin{aligned} ||\mathbf{M}_{\mathbf{Z}_B}\hat{\boldsymbol{D}}||^2 &= ||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{M}_{\mathbf{Z}_B}\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}||^2 \\ \geq &||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0||^2/2 - ||\mathbf{M}_{\mathbf{Z}_B}\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}||^2 \geq n \cdot d_0/2 - ||\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}||^2 \end{aligned}$$

Since $\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}$ follows $\sigma_2^2 \cdot \chi_{|A|}^2$, $P(||\mathbf{M}_{\mathbf{Z}_B}\hat{\boldsymbol{D}}||^2 > 0) \to 1$. When $||\mathbf{M}_{\mathbf{Z}_B}\hat{\boldsymbol{D}}||^2 > 0$, denote $\mathbf{X} = (\hat{\boldsymbol{D}}, \mathbf{Z}_B)$, $\mathbf{X}$ has full column rank. Rewrite $\boldsymbol{Y} = \beta^0\hat{\boldsymbol{D}} + \mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}$. Then, we obtain the unique oracle estimator:

$$(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}_B^{or})^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y} = (\beta^0, \boldsymbol{\alpha}_B^0)^T + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right).$$

Now we show $\frac{\mathbf{X}^T\mathbf{X}}{n} \xrightarrow{p} \mathbf{\Sigma}$. Denote $\boldsymbol{e} = (\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi}$, we have

$$\frac{\mathbf{X}^T\mathbf{X}}{n} = \begin{pmatrix} (\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e})^T(\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e}) & (\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e})^T\mathbf{Z}_B \\ \mathbf{Z}_B^T(\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e}) & \mathbf{Z}_B^T\mathbf{Z}_B \end{pmatrix} /n.$$

Note that $\boldsymbol{e} \sim N(0, \sigma_2^2(\mathbf{Z}_A^T\mathbf{Z}_A)^{-1})$. Moreover, since $\mathbf{Z}_A^T\mathbf{Z}_A/n \to \mathbf{U}_A$, $\boldsymbol{e} \xrightarrow{p} \mathbf{0}_{|A|\times 1}$. Thus,

$$\begin{aligned} \frac{\boldsymbol{e}^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{e}}{n} &\xrightarrow{p} \mathbf{0}_{1\times|A|}\mathbf{U}_A\mathbf{0}_{|A|\times 1} = 0, \\ \frac{(\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{e}}{n} &\xrightarrow{p} (\boldsymbol{\gamma}_A^0)^T\mathbf{U}_A\mathbf{0}_{|A|\times 1} = 0, \\ \frac{\mathbf{Z}_B^T\mathbf{Z}_A\boldsymbol{e}}{n} &\xrightarrow{p} \mathbf{U}_{AB}^T\mathbf{0}_{|A|\times 1} = \mathbf{0}_{|B|\times 1}, \end{aligned}$$

implying that $\frac{\mathbf{X}^T\mathbf{X}}{n} \xrightarrow{p} \mathbf{\Sigma}$.

Note that

$$\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon} \sim N\left(0, (\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)\mathbf{I} - (2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)\mathbf{P}_{\mathbf{Z}_A}\right),$$

so

$$\begin{aligned} &\frac{\mathbf{X}_0^T\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)}{\sqrt{n}} \xrightarrow{d} N\left(0, \mathbf{\Pi}\right), \\ &\mathbf{\Pi} = \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right)\mathbf{\Sigma} - \left(2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right)\mathbf{\Psi}. \end{aligned}$$

Now, $\hat{\boldsymbol{D}} = \mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}$. Since $|\frac{\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\epsilon}}{\sqrt{n}}| \le ||\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}|| \cdot ||\frac{\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\epsilon}}{\sqrt{n}}||$, and $\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi} \sim \sigma_2^2 \cdot \chi^2_{|A|}$, $||\frac{\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\epsilon}}{\sqrt{n}}|| \xrightarrow{p} 0$, we get $\frac{\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\epsilon}}{\sqrt{n}} \xrightarrow{p} 0$. Note that $\frac{\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi})}{\sqrt{n}} = 0$. Then,

$$\begin{aligned}&\frac{\mathbf{X}^T(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi}+\boldsymbol{\epsilon})}{\sqrt{n}}\\ =&\frac{\mathbf{X}_0^T(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi}+\boldsymbol{\epsilon})}{\sqrt{n}}+\frac{(\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi},\mathbf{0}_{n\times|B|})^T(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi}+\boldsymbol{\epsilon})}{\sqrt{n}}\xrightarrow{d}N\left(0,\boldsymbol{\Pi}\right).\end{aligned}\tag{A.1}$$

So,

$$\begin{aligned}&\sqrt{n}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi}+\boldsymbol{\epsilon})=(\frac{\mathbf{X}^T\mathbf{X}}{n})^{-1}\frac{\mathbf{X}^T(\beta^0(\mathbf{I}-\mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi}+\boldsymbol{\epsilon})}{\sqrt{n}}\\ &\xrightarrow{d}\boldsymbol{\Sigma}^{-1}N(0,\boldsymbol{\Pi})=N(0,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}\boldsymbol{\Sigma}^{-1}).\end{aligned}$$

Thus, $\sqrt{n}(\hat{\beta}^{or}-\beta^0)\xrightarrow{d}N(0,v)$, with

$$v=\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\cdot(\boldsymbol{\Sigma}^{-1})_{11}-\left(2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\cdot\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)_{11}.$$

Under the null hypothesis $\beta^0=0$, $\sigma^2=\sigma_1^2\cdot(\boldsymbol{\Sigma}^{-1})_{11}$. □

It is somewhat surprising that the asymptotic variance v depends on β^0 in view of the results of Zhu et al. [4] and Windmeijer et al. [3], in which the asymptotic variances are independent of β^0. This is because Proposition 2 of Zhu et al. [4] assumes that the error term $\boldsymbol{\epsilon}$ is independent of the covariates, whereas Windmeijer et al. [3] only considers relevant IVs without invalid IVs in their model, i.e. their set A, must be a subset of all IVs in their model, although it has similar model setups as ours. In particular, in (5) of Windmeijer et al. [3],

$$\boldsymbol{Y}=\beta^0\cdot\hat{\boldsymbol{D}}+\mathbf{Z}_A\boldsymbol{\alpha}_A^0+\boldsymbol{\xi},\tag{A.2}$$

where $\boldsymbol{\xi}$ is defined implicitly. Actually, since their true model is

$$\begin{aligned}\boldsymbol{D} &= \mathbf{Z}\boldsymbol{\gamma}^0 + \boldsymbol{v},\\ \boldsymbol{Y} &= \beta^0 \cdot \boldsymbol{D} + \mathbf{Z}_A\boldsymbol{\alpha}_A^0 + \boldsymbol{\epsilon},\end{aligned}$$

and

$$\hat{\boldsymbol{D}} = \mathbf{P}_{\mathbf{Z}}\boldsymbol{D} = \mathbf{Z}\boldsymbol{\gamma}^0 + \mathbf{P}_{\mathbf{Z}}\boldsymbol{v}.$$

We can rewrite

$$\boldsymbol{Y} = \beta^0 \cdot \hat{\boldsymbol{D}} + \mathbf{Z}_A\boldsymbol{\alpha}_A^0 + \boldsymbol{\epsilon} + \beta^0(\mathbf{I} - \mathbf{P}_Z)\boldsymbol{v},$$

so explicitly

$$\boldsymbol{\xi} = \boldsymbol{\epsilon} + \beta^0(\mathbf{I} - \mathbf{P}_Z)\boldsymbol{v}. \tag{A.3}$$

Their oracle estimator is

$$\hat{\beta}^{or} = \left(\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{Y}. \tag{A.4}$$

Plugging (A.2) into (A.4), we get

$$\hat{\beta}^{or} = \beta^0 + \left(\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi}. \tag{A.5}$$

Similarly, plugging (A.3) into (A.5) yields that

$$\begin{aligned}\hat{\beta}^{or} =& \beta^0 + \left(\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\epsilon} +\\ & \beta^0 \cdot \left(\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^T\mathbf{M}_{\mathbf{Z}_A}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\boldsymbol{v}.\end{aligned}$$

However, since $\mathbf{Z}_A$ is a subset of $\mathbf{Z}$ and $\mathbf{I} - \mathbf{P}_{\mathbf{Z}} = \mathbf{M}_{\mathbf{Z}}$,

$$\mathbf{M}_{\mathbf{Z}_A}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) = \mathbf{I} - \mathbf{P}_{\mathbf{Z}}. \tag{A.6}$$

Note that $\hat{\boldsymbol{D}} = \mathbf{P_Z}\boldsymbol{D}$ and $\mathbf{P_Z}(\mathbf{I} - \mathbf{P_Z}) = 0$. Then,

$$\begin{aligned}
&\hat{\boldsymbol{D}}^T \mathbf{M}_{\mathbf{Z}_A}(\mathbf{I} - \mathbf{P_Z})\boldsymbol{v} \\
=&\boldsymbol{D}^T \mathbf{P_Z} \mathbf{M}_{\mathbf{Z}_A}(\mathbf{I} - \mathbf{P_Z})\boldsymbol{v} \\
=&\boldsymbol{D}^T \mathbf{P_Z}(\mathbf{I} - \mathbf{P_Z})\boldsymbol{v} = \mathbf{0},
\end{aligned}$$

where the key is that $\mathbf{Z}_A$ is a subset of $\mathbf{Z}$ in (A.6).

In contrast, in our model, the set of invalid IVs might not be a subset of the set of relevant IVs, so (A.6) does not hold. This explains why our asymptotic variance v depends on β. However, if the subset of invalid IVs (i.e. B) is a subset of all relevant IVs (i.e. A), then in Theorem 2 yields that $\boldsymbol{\Sigma} = \boldsymbol{\Psi}$ and our asymptotic variance v is independent of β^0.

Next, we present a lemma to be used in the proof of Theorem 3.

Lemma 1. *For two vectors $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^n$, if $\frac{||\boldsymbol{A}||}{||\boldsymbol{B}||} = k > 1$, then the largest absolute eigenvalue of $\mathbf{P}_{\boldsymbol{A}+\boldsymbol{B}} - \mathbf{P}_{\boldsymbol{A}}$ is no greater than $\frac{1}{k}$.*

Proof. Let $\theta = \frac{\boldsymbol{A}^T \cdot \boldsymbol{B}}{||\boldsymbol{A}|| \cdot ||\boldsymbol{B}||}$. Then $|\theta| \leq 1$. Define $\boldsymbol{e}_1 = \frac{\boldsymbol{A}}{||\boldsymbol{A}||}$, $\boldsymbol{u}_2 = \boldsymbol{B} - \frac{\theta}{k}\boldsymbol{A}$, and $\boldsymbol{e}_2 = \frac{\boldsymbol{u}_2}{||\boldsymbol{u}_2||}$. Hence, $||\boldsymbol{e}_1|| = ||\boldsymbol{e}_2|| = 1$ and $\boldsymbol{e}_1^T \cdot \boldsymbol{e}_2 = 0$. Following the Gram-Schmidt procedure, we generate $\boldsymbol{e}_3, \cdots, \boldsymbol{e}_n$, such that $||\boldsymbol{e}_i|| = 1, i = 1, \cdots, n$ and $\boldsymbol{e}_i^T \cdot \boldsymbol{e}_j = 0$, for $i \neq j, 1 \leq i, j \leq n$. Let $\mathbf{Q} = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n) \in \mathbb{R}^{n \times n}$, $\mathbf{Q}$ is orthonormal. Then,

$$\boldsymbol{A} = \mathbf{Q} \cdot \begin{pmatrix} ||\boldsymbol{A}|| \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boldsymbol{A} + \boldsymbol{B} = \mathbf{Q} \cdot \begin{pmatrix} ||\boldsymbol{A}||(1 + \frac{\theta}{k}) \\ ||\boldsymbol{u}_2|| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Therefore,

$$\mathbf{P}_{\boldsymbol{A}} = \frac{\boldsymbol{A}\cdot\boldsymbol{A}^T}{\boldsymbol{A}^T\cdot\boldsymbol{A}} = \mathbf{Q}\cdot\begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}\cdot\mathbf{Q}^T,$$

$$\mathbf{P}_{\boldsymbol{A}+\boldsymbol{B}} = \mathbf{Q}\cdot\begin{pmatrix} \frac{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & \frac{||\boldsymbol{A}||\cdot||\boldsymbol{u}_2||(1+\frac{\theta}{k})}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & 0 & \cdots & 0 \\ \frac{||\boldsymbol{A}||\cdot||\boldsymbol{u}_2||(1+\frac{\theta}{k})}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & \frac{||\boldsymbol{u}_2||^2}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}\cdot\mathbf{Q}^T.$$

So the non-zero eigenvalues of $\mathbf{P}_{\boldsymbol{A}+\boldsymbol{B}} - \mathbf{P}_{\boldsymbol{A}}$ are those of

$$\mathbf{D} = \begin{pmatrix} -\frac{||\boldsymbol{u}_2||^2}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & \frac{||\boldsymbol{A}||\cdot||\boldsymbol{u}_2||(1+\frac{\theta}{k})}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} \\ \frac{||\boldsymbol{A}||\cdot||\boldsymbol{u}_2||(1+\frac{\theta}{k})}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} & \frac{||\boldsymbol{u}_2||^2}{||\boldsymbol{A}||^2(1+\frac{\theta}{k})^2+||\boldsymbol{u}_2||^2} \end{pmatrix} = \begin{pmatrix} -\frac{1-\theta^2}{k^2(1+\frac{\theta}{k})^2+1-\theta^2} & \frac{k\sqrt{1-\theta^2}(1+\frac{\theta}{k})}{k^2(1+\frac{\theta}{k})^2+1-\theta^2} \\ \frac{k\sqrt{1-\theta^2}(1+\frac{\theta}{k})}{k^2(1+\frac{\theta}{k})^2+1-\theta^2} & \frac{1-\theta^2}{k^2(1+\frac{\theta}{k})^2+1-\theta^2} \end{pmatrix}$$

Easily, the eigenvalues of $\mathbf{D}$ are $\pm\sqrt{\frac{1-\theta^2}{k^2+2k\theta+1}}$. When $\theta = -\frac{1}{k}$, $\frac{1-\theta^2}{k^2+2k\theta+1}$ reaches its maximum $\frac{1}{k^2}$. So, the maximum absolute value of eigenvalues of $\mathbf{P}_{\boldsymbol{A}+\boldsymbol{B}} - \mathbf{P}_{\boldsymbol{A}}$ is no greater than $\frac{1}{k}$. □

Theorem 3. *Under Assumptions 1 to 4, if $K_1 = |A|$ and $K_2 = p_0$, then $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$, either as $n \to \infty$ with a fixed p, or as $n, p \to \infty$.*

Proof. By Assumption 1, it follows from Theorem 2 that the oracle estimator is unique. By Assumptions 2 and 3, it follows from Theorem 3 of [2] that $P(\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}^{or}) \to 1$. Now we focus on the situation of $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}^{or}$ and $\hat{\boldsymbol{D}} = \mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}$. Let

$S(\beta, \boldsymbol{\alpha}) = ||\boldsymbol{Y} - \beta \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\boldsymbol{\alpha}||^2$. Then,

$$\begin{aligned}
I &= P((\hat{\beta}, \hat{\boldsymbol{\alpha}}) \neq (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \\
&\leq P(\min_{\{\boldsymbol{\alpha}|\frac{1}{\tau}\sum_{j=1}^{p}\min(|\alpha_j|,\tau)\leq K_2\}\setminus\{\boldsymbol{\alpha}|supp(\boldsymbol{\alpha})=B\}} S(\beta, \boldsymbol{\alpha}) \leq S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \\
&\leq \sum_{B_1\in\mathbb{B}} P(\min_{\{\boldsymbol{\alpha}\big||\boldsymbol{\alpha}_{B_1}|>\tau, |\boldsymbol{\alpha}_{B_1^c}|\leq\tau\}} S(\beta, \boldsymbol{\alpha}) \leq S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \\
&\leq \sum_{B_1\in\mathbb{B}} P(\min_{\{\boldsymbol{\alpha}\big||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\}} S(\beta, \boldsymbol{\alpha}) \leq S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})).
\end{aligned}$$

For each B_1, we bound $I_{B_1} = P(\min_{\{\boldsymbol{\alpha}\big||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\}} S(\beta, \boldsymbol{\alpha}) \leq S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or}))$. Towards this end, let $(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) = \text{argmin}_{\{\boldsymbol{\alpha}\big||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\}} S(\beta, \boldsymbol{\alpha})$. For $a = n > 1$,

$$\begin{aligned}
S(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) &\geq \frac{a-1}{a}||\boldsymbol{Y} - \hat{\boldsymbol{D}} \cdot \hat{\beta}^T - \mathbf{Z}_{B_1}\hat{\boldsymbol{\alpha}}_{B_1}^T||^2 - (a-1)||\mathbf{Z}_{B_1^c}\hat{\boldsymbol{\alpha}}_{B_1^c}^T||^2 \\
&\geq \frac{a-1}{a}||(\mathbf{I} - \mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_{B_1}})\boldsymbol{Y}||^2 - (a-1)\cdot p \cdot c_{max}\tau^2.
\end{aligned}$$

Rewrite $\boldsymbol{Y} = \hat{\boldsymbol{D}}\cdot\beta^0 + \mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \beta^0\cdot(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}$. Subsequently, for simplicity, we use notations $\boldsymbol{\epsilon} = \beta^0\cdot(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\cdot\boldsymbol{\xi} + \boldsymbol{\epsilon}$, $\boldsymbol{Y} = \mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}$ since $(\mathbf{I} - \mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_{B_1}})\hat{\boldsymbol{D}} = \mathbf{0}$, where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ and

$$\begin{aligned}
\boldsymbol{\Sigma} =& \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2\right)\mathbf{I} - \left(2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2\right)\mathbf{P}_{\mathbf{Z}_A} \\
=& \mathbf{Q}^T\begin{pmatrix} \sigma_1^2\mathbf{I}_{|A|} & \\ & (\sigma_1^2 + 2\sigma_{12}\beta + \sigma_2^2\beta^2)\mathbf{I}_{n-|A|} \end{pmatrix}\mathbf{Q} = \mathbf{Q}^T\boldsymbol{\Lambda}\mathbf{Q}.
\end{aligned} \tag{A.7}$$

Now,

$$\begin{aligned}
S(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) \geq& \frac{a-1}{a}||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}})\boldsymbol{Y}||^2 - (a-1)\cdot p\cdot c_{max}\tau^2 \\
&+ \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}} - \mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_{B_1}})\boldsymbol{Y},
\end{aligned}$$

and

$$S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = ||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B})\boldsymbol{\epsilon}||^2 + \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B})\boldsymbol{\epsilon}.$$

So,

$$
\begin{aligned}
&S(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) - S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or}) \\
\geq &-\frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0)^T(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0) \\
&+ (a-1)||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2 - \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B})\boldsymbol{\epsilon} - (a-1)\cdot p \cdot c_{max}\tau^2 \\
&+ \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{B_1}})\boldsymbol{Y} - \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_B})\boldsymbol{\epsilon} = -\sum_{l=1}^{4} L_j + \sum_{l=1}^{4} b_j,
\end{aligned}
$$

where $1 > \delta = \delta_1 + \delta_2 + \delta_3$, and $\delta_1, \delta_2, \delta_3 > 0$,

$$
\begin{aligned}
L_1 &= \frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0)^T(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0), \\
L_2 &= \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B})\boldsymbol{\epsilon}, \\
L_3 &= \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\boldsymbol{Y}, \\
L_4 &= \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_B})\boldsymbol{\epsilon}, \\
b_1 &= (a-1-\delta)||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2, \\
b_2 &= \delta_1||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2 - (a-1)\cdot p \cdot c_{max}\tau^2, \\
b_3 &= \delta_2||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2, \\
b_4 &= \delta_3||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2.
\end{aligned}
$$

So, $I_{B_1} \leq P(L_1 \geq b_1) + P(L_2 \geq b_2) + P(L_3 \geq b_3) + P(L_4 \geq b_4)$. For simplicity, we use σ^2 for σ_M^2. Then, $P(L_1 \geq b_1) \leq \frac{E(e^{\frac{t_1 L_1}{\sigma^2}})}{e^{\frac{t_1 b_1}{\sigma^2}}}$. Let $\boldsymbol{v} = \mathbf{Q}(a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0$ and $\mathbf{P} = \mathbf{Q}(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Q}^T$. Then,

$$
\begin{aligned}
P(L_1 \geq b_1) \leq &e^{-\frac{t_1 b_1}{\sigma^2}} \int \exp\left(\frac{t_1}{a\sigma^2}(\boldsymbol{\epsilon} - \boldsymbol{v})^T(\mathbf{I} - \mathbf{P})(\boldsymbol{\epsilon} - \boldsymbol{v})\right) \cdot (2\pi)^{-n/2}|\boldsymbol{\Lambda}|^{-1/2} \cdot \exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\epsilon}}{2}\right) d\boldsymbol{\epsilon}, \\
\leq &e^{-\frac{t_1 b_1}{\sigma^2}} \left(\int \exp\left(\frac{t_1}{a\sigma^2}(\boldsymbol{\epsilon} - \boldsymbol{v})^T(\mathbf{I} - \mathbf{P})(\boldsymbol{\epsilon} - \boldsymbol{v})\right) \cdot (2\pi)^{-n/2}\sigma^{-n} \cdot \exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2\sigma^2}\right) d\boldsymbol{\epsilon}\right) \cdot r^n, \\
\leq &\frac{1}{(1 - 2t_1/a)^{n/2}} \exp\left(-\frac{ni}{d_2\sigma^2} C_{\min 2}\right) \cdot r^n, \\
\leq &\frac{1}{(1 - 2t_1/a)^{n/2}} \exp\left(-\frac{ni}{d_2\sigma^2}(C_{\min 2} - d_2\sigma^2 \log r)\right).
\end{aligned}
$$

Here $i = |B \setminus B_1|$. By Theorem 3 of [2] and Assumption 2, $\sum_{B_1 \in \mathbb{B}} P(L_1 \geq b_1) \to 0$. Similarly, by Assumption 2 and 3, $\sum_{B_1 \in \mathbb{B}} P(L_2 \geq b_2) \to 0$.

To prove that $\sum_{B_1 \in \mathbb{B}} P(L_3 \geq b_3) \to 0$, note that $L_3 = 0$ when $A \subseteq B_1$. So,

$$\sum_{B_1 \in \mathbb{B}} P(L_3 \geq b_3) = \sum_{B_1 \in \mathbb{G}} P(L_3 \geq b_3).$$

Moreover, note that $\mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{\mathbf{B_1}}} = \mathbf{P}_{\mathbf{Z}_{\mathbf{B_1}}} + \mathbf{P}_{\boldsymbol{u}}$ and $\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} = \mathbf{P}_{\mathbf{Z}_{\mathbf{B_1}}} + \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}}$, where $\boldsymbol{u} = \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0$ and $\boldsymbol{v} = \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{P}_{\mathbf{Z}_A} \boldsymbol{\xi}$. So, $\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} - \mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{\mathbf{B_1}}} = \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}$. Hence, for $q = \frac{1}{3}$,

$$\begin{aligned}
&P(L_3 \geq b_3) \\
\leq &P(\boldsymbol{Y}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{Y} \geq b_3) \\
\leq &P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) + P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \boldsymbol{Y}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{Y} \geq b_3).
\end{aligned}$$

By Lemma 1, $P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \boldsymbol{Y}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{Y} \geq b_3) \leq P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3) \leq P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3)$. So

$$P(L_3 \geq b_3) \leq P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) + P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3).$$

Thus, when $B_1 \in \mathbb{B}$, we have

$$\begin{aligned}
&P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3) \\
=&P(\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n^q} \geq b_3) \\
=&P(\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n} \geq \frac{b_3}{n^{1-q}}) \\
\leq &\frac{E(e^{\frac{t}{\sigma^2}\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0+\boldsymbol{\epsilon}||^2}{n}})}{e^{\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}}}
\end{aligned}$$

where

$$
\begin{aligned}
&E(e^{\frac{t}{\sigma^2}\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0+\boldsymbol{\epsilon}||^2}{n}}) \\
&= \int \exp\left(\frac{t}{\sigma^2}\cdot\frac{(\boldsymbol{\epsilon}+\mathbf{Q}\mathbf{Z}_B\boldsymbol{\alpha}_B^0)^T(\boldsymbol{\epsilon}+\mathbf{Q}\mathbf{Z}_B\boldsymbol{\alpha}_B^0)}{n}\right)(2\pi)^{-n/2}|\boldsymbol{\Lambda}|^{-1/2}\exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\epsilon}}{2}\right)d\boldsymbol{\epsilon}.
\end{aligned}
$$

Let $v = \mathbf{Q}\mathbf{Z}_B\boldsymbol{\alpha}_B^0 = (v_1, \cdots, v_n)$. If $\Lambda_i = \sigma_M^2$, then

$$
\int \exp\left(\frac{t}{\sigma^2}\cdot\frac{(\epsilon_i+v_i)^2}{n}\right)(2\pi)^{-1/2}\sigma^{-1}\exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \frac{1}{\sqrt{1-2\frac{t}{n}}}\exp\left(\frac{tv_i^2}{\sigma^2(n-2t)}\right).
$$

If $\Lambda_i = \sigma_m^2$, then,

$$
\begin{aligned}
&\int \exp\left(\tfrac{t}{\sigma^2}\cdot\tfrac{(\epsilon_i+v_i)^2}{n}\right)(2\pi)^{-1/2}\sigma_m^{-1}\exp\left(-\tfrac{\epsilon_i^2}{2\sigma_m^2}\right) \\
&= \tfrac{1}{\sqrt{1-2\frac{t}{nr^2}}}\exp\left(\tfrac{tv_i^2/r^2}{\sigma^2(n-2t/r^2)}\right) \le \tfrac{1}{\sqrt{1-2\frac{t}{n}}}\exp\left(\tfrac{tv_i^2}{\sigma^2(n-2t)}\right). \qquad \text{(A.8)}
\end{aligned}
$$

So,

$$
\begin{aligned}
&P(\frac{1}{n^q}Y^TY \ge b_3) \le \frac{E(e^{\frac{t}{\sigma^2}\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0+\boldsymbol{\epsilon}||^2}{n}})}{e^{\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}}} \\
\le&\frac{(1-2\frac{t}{n})^{-\frac{n}{2}}\cdot e^{\frac{t\cdot||\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{(n-2t)\sigma^2}}}{e^{\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}}} \le C_1\cdot e^{-\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}} \\
=&C_1\cdot e^{-\frac{t}{\sigma^2}\delta_2\cdot n^q i C_{\min 2}}.
\end{aligned}
$$

where C_1 is a positive constant and $i = |B \setminus B_1|$. By similar arguments of Theorem 3 in [2], it follows from Assumption 2 that $\sum_{B_1\in\mathbb{G}} P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \ge b_3) \le \sum_{B_1\in\mathbb{B}} P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \ge b_3) \to 0$ as $n, p \to \infty$.

When $B_1 \in \mathbb{G}$,

$$\begin{aligned}
&P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \\
=&P(\boldsymbol{\xi}^T \mathbf{P}_{\mathbf{Z}_A} \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{P}_{\mathbf{Z}_A} \boldsymbol{\xi} \geq \frac{||\boldsymbol{u}||^2}{n^{2q}}) \\
\leq&P(\boldsymbol{\xi}^T \mathbf{P}_{\mathbf{Z}_A} \boldsymbol{\xi} \geq \frac{||\boldsymbol{u}||^2}{n^{2q}}) \\
\leq&\frac{E(e^{\frac{t}{\sigma_2^2}\boldsymbol{\xi}^T \mathbf{P}_{\mathbf{Z}_A} \boldsymbol{\xi}})}{e^{\frac{t}{\sigma_2^2}\frac{||\boldsymbol{u}||^2}{n^{2q}}}} \\
\leq&C_2 \cdot e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}},
\end{aligned}$$

where C_2 is a positive constant, $i = |A \setminus B_1|$, and $j = |B_1 \setminus A|$. So,

$$\begin{aligned}
&\sum_{B_1 \in \mathbb{G}} P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \\
\leq&C_2 \sum_{B_1 \in \mathbb{G}} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}} \\
\leq&C_2 \sum_{i=1}^{|A|} \sum_{j=0}^{\max(p_0-|A|+i,0)} \binom{|A|}{i} \binom{p-|A|}{j} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}} \\
\leq&C_2 \sum_{i=1}^{|A|} \sum_{j=0}^{\max(p_0-|A|+i,0)} |A|^i (p-|A|)^j e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}} \\
\leq&C_2 \sum_{i=1}^{|A|} |A|^i \frac{(p-|A|)^{\max(p_0-|A|+i,0)+1} - 1}{p-|A|-1} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}} \\
\leq&(2C_2 \sum_{i=1}^{|A|} |A|^i (p-|A|)^i e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 3}}) \cdot p^{p_0} \\
\leq&2C_2 \frac{|A|(p-|A|) e^{-\frac{t}{\sigma_2^2} n^{1-2q} C_{\min 3}}}{1 - |A|(p-|A|) e^{-\frac{t}{\sigma_2^2} n^{1-2q} C_{\min 3}}} \cdot p^{p_0} \\
\leq&C_3 \cdot \exp\{-\frac{t}{\sigma_2^2} n^{\frac{1}{3}} (C_{\min 3} - \frac{1}{t} \frac{(p_0+2)\log p}{n^{\frac{1}{3}}} \cdot \sigma_2^2)\}.
\end{aligned}$$

By Assumption 4, $\sum_{B_1 \in \mathbb{G}} P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \to 0$ as $n, p \to \infty$.

To show $\sum_{B_1\in\mathbb{B}} P(L_4 \geq b_4) \to 0$, note that $\mathbf{P}_{\mathbf{Z}_A\cup\mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_B} = \mathbf{P}_{\mathbf{U}}$, where $\mathbf{U} = \mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A$, so

$$\begin{aligned}
&P(L_4 \geq b_4) \leq P(\boldsymbol{\epsilon}^T\mathbf{P}_{\mathbf{U}}\boldsymbol{\epsilon} \geq b_4)\\
\leq& e^{-\frac{t}{\sigma^2}b_4}\int \exp\left(\frac{t}{\sigma^2}\boldsymbol{\epsilon}^T\mathbf{P}_{\mathbf{U}}\boldsymbol{\epsilon}\right)(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}}{2}\right)d\boldsymbol{\epsilon}\\
\leq& e^{-\frac{t}{\sigma^2}b_4}\left(\int \exp\left(\frac{t}{\sigma^2}\boldsymbol{\epsilon}^T\mathbf{Q}\mathbf{P}_{\mathbf{U}}\mathbf{Q}^T\boldsymbol{\epsilon}\right)(2\pi)^{-n/2}\sigma^{-n}\exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2\sigma^2}\right)d\boldsymbol{\epsilon}\right)r^n.
\end{aligned}$$

Therefore,

$$P(L_4 \geq b_4) \leq C_4\cdot e^{-\frac{t}{\sigma^2}\delta_3 ni(C_{\min 2}-d_2\sigma^2\log r)},$$

where C_4 is some positive constant and $i = |B\setminus B_1|$. So, $\sum_{B_1\in\mathbb{B}} P(L_4 \geq b_4) \to 0$, as $n, p \to \infty$. This establishes that $P((\hat{\beta}, \hat{\boldsymbol{\alpha}}) \neq (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \to 0$, as $n, p \to \infty$. □

Corollary 1. *Assume that Assumptions 1 to 4 are met. If $K_1 = |A|$ and $K_2 = p_0$, then under the null hypothesis of $\beta^0 = 0$, Λ_n converges in distribution to χ^2_1, a chi-squared distribution with degrees of freedom 1, either as $n \to \infty$ with a fixed p, or as both $n, p \to \infty$.*

Proof. By Theorem 3, $P((\hat{\beta}, \hat{\boldsymbol{\alpha}}) \neq (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \to 0$. By Theorem 3 of [2], under the null hypothesis, $P(\hat{\boldsymbol{\alpha}}^{(0)}) \neq \hat{\boldsymbol{\alpha}}_0^{or})) \to 0$, where $\hat{\boldsymbol{\alpha}}_0^{or} = \text{argmin}_{\boldsymbol{\alpha}}\,||\boldsymbol{Y} - \mathbf{Z}_B\boldsymbol{\alpha}_B||^2$.
As in Theorem 2 of [4], we have

$$\Lambda_n = \frac{n\boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_B})\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||^2} + \frac{n\boldsymbol{\epsilon}^T(\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B})\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||^2} + R(\boldsymbol{\epsilon}).$$

Hence, $R(\boldsymbol{\epsilon}) \to 0$ as $n \to \infty$. Then, it suffices to show that $\boldsymbol{\epsilon}^T(\mathbf{P}_{\hat{\boldsymbol{D}}\cup Z_B} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B})\boldsymbol{\epsilon} \to_p 0$.

Note that $\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} = \mathbf{P}_{\mathbf{Z}_B} + \mathbf{P}_{\boldsymbol{u}}$ and $\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B} = \mathbf{P}_{\mathbf{Z}_B} + \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}}$, where $\boldsymbol{u} = \mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0$

and $\boldsymbol{v} = \mathbf{M}_{\mathbf{Z}_B}\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}$. So, $\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} = \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}$. Moreover,

$$\begin{aligned}\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{\epsilon} &= \frac{\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v}} - \frac{\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u}} \\ &= \frac{\boldsymbol{\epsilon}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v}} - \frac{(\boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v})\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v})}.\end{aligned}$$

By Assumption 1, $\frac{\boldsymbol{u}^T\boldsymbol{u}}{n} \geq d_0 > 0$. Note that $\boldsymbol{v}^T\boldsymbol{v}$ follows a χ^2-distribution with the degrees of freedom no greater than $|A|$. Hence, $\frac{\boldsymbol{v}^T\boldsymbol{v}}{n} \to_p 0$. Moreover, $\boldsymbol{u}^T\boldsymbol{v}$ follows $N(0, (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{P}_{\mathbf{Z}_A}\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_2^2)$, so $\frac{\boldsymbol{u}^T\boldsymbol{v}}{n} \to_p 0$. Again, $\boldsymbol{\epsilon}^T\boldsymbol{u}$ follows $N(0, (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_1^2)$. So $\frac{\boldsymbol{\epsilon}^T\boldsymbol{u}}{\sqrt{n}}$ has the variance no greater than $\frac{(\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_1^2}{n}$. Since $\boldsymbol{\epsilon}^T\boldsymbol{v} = \boldsymbol{\epsilon}^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A(\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi}$, $(\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi}$ follows $N(0, (\mathbf{Z}_A^T\mathbf{Z}_A)^{-1})$. As a result, $(\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi} \to_p 0$. Finally, $\boldsymbol{\epsilon}^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A$ follows $N(0, \mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A)$. So, $\frac{\boldsymbol{\epsilon}^T\boldsymbol{v}}{\sqrt{n}} \to_p 0$. Consequently, $\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{\epsilon} \to_p 0$ as $n \to \infty$. Thus $\Lambda_n \to_d \chi_1^2$. □

Theorem 4. *Assume that Assumptions 1 to 4 are met. Then when p is fixed, if $|A| \in \mathcal{K}_1$ and $p_0 \in \mathcal{K}_2$, we have $P\left(\hat{K}_1 = |A|, \hat{K}_2 = p_0\right) \to 1$ as $n \to \infty$.*

Proof. To be clear, in the following of the proof we denote $|A| = p_1$ and $|B| = p_0 = p_2$. First we show $P\left(\hat{K}_1 = p_1\right) \to 1$, which is equivalent to $P\left(\mathrm{BIC}_1(K_1) \leq \mathrm{BIC}_1(p_1)\right) \to 0$ for any $K_1 \neq p_1$. From Theorem 3, we have $P(\hat{\boldsymbol{\gamma}}_{p_1} = \hat{\boldsymbol{\gamma}}^{or}) \to 1$, so $P(||\hat{\boldsymbol{\gamma}}_{p_1}||_0 = p_1) \to 1$. So we have

$$\begin{aligned}\mathrm{BIC}_1(p_1) &= n \cdot \log\frac{||\boldsymbol{D} - \mathbf{Z}\hat{\boldsymbol{\gamma}}_{p_1}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\gamma}}_{p_1}||_0 \\ &\leq n \cdot \log\frac{||\boldsymbol{D} - \mathbf{Z}\boldsymbol{\gamma}^0||^2}{n} + \log(n) \cdot p_1 \\ &= n \cdot \log\frac{||\boldsymbol{\xi}||^2}{n} + \log(n) \cdot p_1.\end{aligned}$$

Suppose for some $A_1 \subseteq S$, $|A_1| = K_1$, and for $i \in A_1^c$, $|\left(\hat{\boldsymbol{\gamma}}_{K_1}\right)_i| \leq \tau_1$. With Assumption

3, following similar argument in proof of Theorem 3, we have

$$\begin{aligned}\mathrm{BIC}_1(K_1) &= n \cdot \log \frac{||\boldsymbol{D} - \mathbf{Z}\hat{\boldsymbol{\gamma}}_{K_1}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\gamma}}_{K_1}||_0 \\ &\geq n \cdot \log \frac{||\boldsymbol{D} - \mathbf{Z}_{A_1}(\hat{\boldsymbol{\gamma}}_{K_1})_{A_1}||^2 - n \cdot p \cdot c_{max} \cdot \tau_1^2}{n} + \log(n) \cdot K_1 \\ &\geq n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{D}||^2 - 6\sigma_2^2}{n} + \log(n) \cdot K_1.\end{aligned}$$

So we have

$$\begin{aligned}&P\left(\mathrm{BIC}_1(K_1) \leq \mathrm{BIC}_1(p_1)\right) \\ \leq &P\left(n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{D}||^2 - 6\sigma_2^2}{n} + \log(n) \cdot K_1 \leq n \cdot \log \frac{||\boldsymbol{\xi}||^2}{n} + \log(n) \cdot p_1\right).\end{aligned}$$

We have

$$\begin{aligned}\frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{D}||^2 - 6\sigma_2^2}{n} &= \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \left(\mathbf{Z}_A \boldsymbol{\gamma}_A^0 + \boldsymbol{\xi}\right)||^2 - 6\sigma_2^2}{n} \\ = &\frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0||^2}{n} + \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{\xi}||^2}{n} + 2 \cdot \frac{\boldsymbol{\xi}^T \mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0 - 6\sigma_2^2}{n}.\end{aligned}$$

The last term

$$\frac{\boldsymbol{\xi}^T \mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0 - 6\sigma_2^2}{n} \sim N\left(\frac{-6\sigma_2^2}{n}, \sigma_2^2 \cdot \frac{(\mathbf{Z}_A \boldsymbol{\gamma}_A^0)^T \mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0}{n^2}\right) = o_p(1),$$

then

$$\begin{aligned}&P\left(\mathrm{BIC}_1(K_1) \leq \mathrm{BIC}_1(p_1)\right) \\ \leq &P\left(n \cdot \log\left(\frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0||^2}{n} + \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{\xi}||^2}{n} + o_p(1)\right) + \log(n) \cdot K_1 \leq n \cdot \log \frac{||\boldsymbol{\xi}||^2}{n} + \log(n) \cdot p_1\right).\end{aligned}$$

Here

$$\frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{\xi}||^2}{n} \sim \sigma_2^2 \cdot \frac{\chi^2_{n-K_1}}{n} \to \sigma_2^2, \quad \frac{||\boldsymbol{\xi}||^2}{n} \sim \sigma_2^2 \cdot \frac{\chi^2_n}{n} \to \sigma_2^2.$$

When $K_1 < p_1$, with Assumption 2, we have

$$\liminf_{n\to\infty} \frac{||\mathbf{M}_{\mathbf{Z}_{A_1}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0||^2}{n} = C,$$

for some positive constant C, and

$$P\left(\mathrm{BIC}_1(K_1) \leq \mathrm{BIC}_1(p_1)\right) = P\left(n \cdot \log\left(1 + \frac{C}{\sigma_2^2} + o_p(1)\right) \leq \log(n) \cdot (p_1 - K_1)\right) \to 0. \tag{A.9}$$

When $K_1 > p_1$, we have $||\mathbf{M}_{\mathbf{Z}_{A_1}} \boldsymbol{D}||^2$ follows $\sigma_2^2 \chi^2_{n-K_1}$, and

$$P\left(\mathrm{BIC}_1(K_1) \leq \mathrm{BIC}_1(p_1)\right) = P\left(\log(n) \cdot (K_1 - p_1) \leq n \cdot \log(1 + \frac{\chi^2_{K_1} + 6\sigma_2^2}{\chi^2_{n-K_1} - 6\sigma_2^2})\right) \to 0. \tag{A.10}$$

Combining A.9 and A.10, we get $P\left(\hat{K}_1 = p_1\right) \to 1$ as $n \to \infty$.

Now we show $P\left(\hat{K}_2 = p_2\right) \to 1$. With $P\left(\hat{K}_1 = p_1\right) \to 1$, we only need to consider the case $\hat{A} = A$, thus $\hat{\boldsymbol{D}} = \mathbf{Z}_A \boldsymbol{\gamma}_A^0 + \mathbf{P}_{\mathbf{Z}_A} \boldsymbol{\xi}$. From Theorem 3, we have $P(\hat{\boldsymbol{\alpha}}_{p_2} = \hat{\boldsymbol{\alpha}}^{or}) \to 1$, so $P(||\hat{\boldsymbol{\alpha}}_{p_2}||_0 = p_2) \to 1$. We have

$$\begin{aligned}
\mathrm{BIC}_2(p_2) &= n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{p_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{p_2}||^2}{n} + \log(n) \cdot p_2 \\
&\leq n \cdot \log \frac{||\boldsymbol{Y} - \beta^0 \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\boldsymbol{\alpha}^0||^2}{n} + \log(n) \cdot p_2 \\
&= n \cdot \log \frac{||\beta^0 \cdot \mathbf{M}_{\mathbf{Z}_A} \boldsymbol{\xi} + \boldsymbol{\epsilon}||^2}{n} + \log(n) \cdot p_2.
\end{aligned}$$

Suppose for some $B_1 \subseteq S$, $|B_1| = K_2$ and $|\,(\hat{\boldsymbol{\alpha}}_{K_2})_{B_1^c}\,| \leq \tau_2$. With Assumption 3,

following similar argument in proof of Theorem 3, we have

$$\begin{aligned}\mathrm{BIC}_2(K_2) &= n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{K_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{K_2}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\alpha}}_{K_2}||_0 \\ &\geq n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{K_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}_{B_1}(\hat{\boldsymbol{\alpha}}_{K_2})_{B_1}||^2 - n \cdot p \cdot c_{max}\tau_2^2}{n} + \log(n) \cdot K_2 \\ &\geq n \cdot \log \frac{||\mathbf{M}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2.\end{aligned}$$

So we have

$$\begin{aligned}&P\left(\mathrm{BIC}_2(K_2) \leq \mathrm{BIC}_2(p_2)\right) \\ \leq &P\left(n \cdot \log \frac{||\mathbf{M}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \leq n \cdot \log \frac{||\beta^0 \cdot \mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi} + \boldsymbol{\epsilon}||^2}{n} + \log(n) \cdot p_2\right).\end{aligned}$$

With Assumption 4 and similar arguments in proof of Theorem 3, we get

$$\begin{aligned}&P\left(\mathrm{BIC}_2(K_2) \leq \mathrm{BIC}_2(p_2)\right) \\ \leq &\, P(n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \leq n \cdot \log \frac{||\beta^0 \cdot \mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi} + \boldsymbol{\epsilon}||^2}{n} + \log(n) \cdot p_2) \\ = &\, P(n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} (\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \\ &\leq n \cdot \log \frac{||\beta^0 \cdot \mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi} + \boldsymbol{\epsilon}||^2}{n} + \log(n) \cdot p_2).\end{aligned}$$

We have

$$\begin{aligned}&\frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \left(\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}\right)||^2 - 6\sigma_M^2}{n} \\ = &\frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B \boldsymbol{\alpha}_B^0||^2}{n} + \frac{||\mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \left(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}\right)||^2}{n} + 2 \cdot \frac{\left(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}\right)^T \mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B \boldsymbol{\alpha}_B^0 - 6\sigma_M^2}{n},\end{aligned}$$

and the last term

$$\begin{aligned}&\frac{\left(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}\right)^T \mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B \boldsymbol{\alpha}_B^0 - 6\sigma_M^2}{n} \\ \sim &N\left(\frac{-6\sigma_M^2}{n}, \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right) \frac{(\boldsymbol{\alpha}_B^0)^T \mathbf{Z}_B^T \mathbf{M}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B \boldsymbol{\alpha}_B^0}{n^2}\right) = o_p(1),\end{aligned}$$

so we have

$$\begin{aligned}
&P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_2)\right) \\
\leq\ & P(n\log(\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n} + \frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\left(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}\right)||^2}{n} + o_p(1)) \\
&+\log(n)\cdot(K_2-p_2) \leq n\log\frac{||\beta^0\cdot\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi}+\boldsymbol{\epsilon}||^2}{n}),
\end{aligned}$$

here

$$\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\left(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}\right)||^2}{n} \sim \left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\frac{\chi^2_{n-K_2-1}}{n} \to \left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right),$$

$$\begin{aligned}
\frac{||\beta^0\cdot\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi}+\boldsymbol{\epsilon}||^2}{n} &\sim \frac{||N\left(\mathbf{0},(\sigma_1^2+2\sigma_{12}\beta^0+\sigma_2^2(\beta^0)^2)\,\mathbf{I}-(2\sigma_{12}\beta^0+\sigma_2^2(\beta^0)^2)\,\mathbf{P}_{\mathbf{Z}_A}\right)||^2}{n} \\
&\to \left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right).
\end{aligned}$$

When $K_2 < p_2$, with Assumption 2, we have

$$\liminf_{n\to\infty}\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n} = C,$$

for some positive constant C, thus

$$\begin{aligned}
&P\left(\mathrm{BIC}_2(K_2) \leq \mathrm{BIC}_2(p_2)\right) \\
=&P\left(n\cdot\log(1+\frac{C}{\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2}+o_p(1)) \leq \log(n)\cdot(p_2-K_2)\right) \to 0.
\end{aligned} \tag{A.11}$$

When $K_2 > p_2$,

$$
\begin{aligned}
&P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_2)\right) \\
=&P\left(\log(n)\cdot(K_2-p_2) \leq n\cdot\log\frac{||\beta^0\cdot\mathbf{M}_{\mathbf{Z}_A}\boldsymbol{\xi}+\boldsymbol{\epsilon}||^2}{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi})||^2-6\sigma_M^2}\right) \\
=&P\left(\log(n)\cdot(K_2-p_2) \leq n\cdot\log\frac{||\beta^0\cdot\boldsymbol{\xi}+\boldsymbol{\epsilon}||^2+||\beta^0\cdot\mathbf{P}_{\mathbf{Z}_A}\boldsymbol{\xi}||^2+2\beta^0\cdot\boldsymbol{\xi}^T\mathbf{P}_{\mathbf{Z}_A}(\beta^0\cdot\boldsymbol{\xi}+\boldsymbol{\epsilon})}{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi})||^2-6\sigma_M^2}\right) \quad \text{(A.12)} \\
=&P\left(\log(n)\cdot(K_2-p_2) \leq n\cdot\log(\frac{\chi^2_n+O_p(1)}{\chi^2_{n-K_2-1}-6\sigma_M^2})\right) \\
=&P\left(\log(n)\cdot(K_2-p_2) \leq n\cdot\log(1+\frac{\chi^2_{K_2+1}+O_p(1)}{\chi^2_{n-K_2-1}-6\sigma_M^2})\right)\to 0.
\end{aligned}
$$

Combining (A.11) and (A.12) we get $P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_2)\right) \to 0$ for any $K_2 \neq p_2$, thus $P\left(\hat{K}_2 = p_2\right) \to 1$ as $n \to \infty$. □

Theorem 5. *With Assumptions 1 and 5 satisfied, the probability of the oracle estimator $\hat{\beta}^{or}$ defined in (9) being unique converges to 1 as $n, n_2 \to \infty$, and $\hat{\beta}^{or}$ is a consistent estimator of true causal effect β^0 with $\hat{\beta}^{or} \xrightarrow{p} \beta^0$ as $n, n_2 \to \infty$. Furthermore, we have $\sqrt{n}(\hat{\beta}^{or} - \beta^0) \xrightarrow{d} N(0, v)$, with variance $v = (\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)\cdot(\boldsymbol{\Sigma}^{-1})_{11} + w(\beta^0)^2\sigma_2^2(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_2\boldsymbol{\Sigma}^{-1})_{11}$. Under the null hypothesis H_0: $\beta^0 = 0$, we have $v = \sigma_1^2\cdot(\boldsymbol{\Sigma}^{-1})_{11}$.*

Proof. Note that

$$||\mathbf{M}_{\mathbf{Z}_B}\hat{\boldsymbol{D}}||^2 = ||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{e}||^2 \geq ||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0||^2/2 - ||\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{e}||^2 \geq n\cdot d_0/2 - ||\mathbf{Z}_A\boldsymbol{e}||^2.$$

Then, $||\mathbf{Z}_A\boldsymbol{e}||^2/n = \boldsymbol{e}^T(\mathbf{Z}_A^T\mathbf{Z}_A/n)\boldsymbol{e}$, since $\boldsymbol{e} \xrightarrow{p} \mathbf{0}_{|A|\times 1}$, $\mathbf{Z}_A^T\mathbf{Z}_A/n \to \mathbf{U}_A$ as $n, n_2 \to \infty$. Hence, $P(||\mathbf{M}_{\mathbf{Z}_B}\hat{\boldsymbol{D}}||^2 > 0) \to 1$, which means the oracle estimator is unique with probability tending to 1 as $n \to \infty$. Rewrite

$$\boldsymbol{Y} = \beta^0\cdot\hat{\boldsymbol{D}} + \mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon} + \beta^0\cdot\boldsymbol{\xi} - \beta^0\cdot\mathbf{Z}_A\boldsymbol{e}.$$

Let $\mathbf{X} = (\hat{\boldsymbol{D}}, \mathbf{Z}_B)$ and $\mathbf{X}_0 = (\mathbf{Z}_A\boldsymbol{\gamma}_A^0, \mathbf{Z}_B)$. Now,

$$\begin{pmatrix} \hat{\beta}^{or} \\ \hat{\boldsymbol{\alpha}}_B^{or} \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y} = \begin{pmatrix} \beta^0 \\ \boldsymbol{\alpha}_B^0 \end{pmatrix} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\epsilon} + \beta^0 \cdot \boldsymbol{\xi} - \beta^0 \cdot \mathbf{Z}_A\boldsymbol{e}).$$

So,

$$\sqrt{n}\left(\begin{pmatrix} \hat{\beta}^{or} \\ \hat{\boldsymbol{\alpha}}_B^{or} \end{pmatrix} - \begin{pmatrix} \beta^0 \\ \boldsymbol{\alpha}_B^0 \end{pmatrix}\right) = \sqrt{n}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\epsilon} + \beta^0 \cdot \boldsymbol{\xi} - \beta^0 \cdot \mathbf{Z}_A\boldsymbol{e})$$
$$= \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1} \frac{\mathbf{X}^T(\boldsymbol{\epsilon} + \beta^0 \cdot \boldsymbol{\xi} - \beta^0 \cdot \mathbf{Z}_A\boldsymbol{e})}{\sqrt{n}}.$$

Note that

$$\frac{\mathbf{X}^T\mathbf{X}}{n} = \begin{pmatrix} (\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e})^T(\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e}) & (\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e})^T\mathbf{Z}_B \\ \mathbf{Z}_B^T(\mathbf{Z}_A\boldsymbol{\gamma}_A^0 + \mathbf{Z}_A\boldsymbol{e}) & \mathbf{Z}_B^T\mathbf{Z}_B \end{pmatrix} /n.$$

Moreover, $\boldsymbol{e} \xrightarrow{p} \mathbf{0}_{|A|\times 1}$ as $n_2 \to \infty$. As $n, n_2 \to \infty$,

$$\frac{\boldsymbol{e}^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{e}}{n} \xrightarrow{p} 0, \quad \frac{(\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{e}}{n} \xrightarrow{p} 0, \quad \frac{\mathbf{Z}_B^T\mathbf{Z}_A\boldsymbol{e}}{n} \xrightarrow{p} 0.$$

So,

$$\frac{\mathbf{X}^T\mathbf{X}}{n} \xrightarrow{p} \boldsymbol{\Sigma}. \tag{A.13}$$

Thus,

$$\frac{\mathbf{Z}_A^T}{\sqrt{n}}(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}) \sim N\left(0, (\sigma_1^2 + 2\beta^0\sigma_{12} + (\beta^0)^2\sigma_2^2)\frac{\mathbf{Z}_A^T\mathbf{Z}_A}{n}\right),$$

which, together with the fact that $\boldsymbol{e} \xrightarrow{p} \mathbf{0}_{|A|\times 1}$, yields that

$$\frac{\boldsymbol{e}^T\mathbf{Z}_A^T(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})}{\sqrt{n}} \xrightarrow{p} 0. \tag{A.14}$$

Note that $\boldsymbol{\Theta} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, $\mathbf{Q}$ is orthonormal and $\boldsymbol{\Lambda}$ is diagonal. Let $\boldsymbol{e}^* = \mathbf{Q}^T\boldsymbol{e} \sim$

$N(0, \mathbf{\Lambda})$. Then

$$\frac{\boldsymbol{e}^T \mathbf{Z}_A^T \mathbf{Z}_A \boldsymbol{e}}{\sqrt{n}} = \frac{(\boldsymbol{e}^*)^T \mathbf{Q}^T \mathbf{Z}_A^T \mathbf{Z}_A \mathbf{Q} \boldsymbol{e}^*}{\sqrt{n}}.$$

Note that $\frac{\mathbf{Z}_A^T \mathbf{Z}_A}{n} \to \mathbf{U}_A$. Then as the eigenvalues of $\frac{\mathbf{Z}_A^T \mathbf{Z}_A}{n}$ are upper-bounded by u_1,

$$\frac{(\boldsymbol{e}^*)^T \mathbf{Q}^T \mathbf{Z}_A^T \mathbf{Z}_A \mathbf{Q} \boldsymbol{e}^*}{\sqrt{n}} \leq \sqrt{n} u_1 (\boldsymbol{e}^*)^T \boldsymbol{e}^*.$$

Note that $n_2 \mathbf{\Theta} \to \mathbf{\Theta}_0$. Then the eigenvalues of $\mathbf{\Lambda}$ are upper-bounded by $\frac{u_2}{n_2}$. Thus,

$$\sqrt{n} u_1 (\boldsymbol{e}^*)^T \boldsymbol{e}^* \leq u_1 \cdot u_2 \cdot \frac{\sqrt{n}}{n_2} \chi^2_{|A|}.$$

As $\frac{\sqrt{n}}{n_2} \to 0$, we have

$$\frac{\boldsymbol{e}^T \mathbf{Z}_A^T \mathbf{Z}_A \boldsymbol{e}}{\sqrt{n}} \xrightarrow{p} 0. \tag{A.15}$$

A combination of (A.14) and (A.15) yields that

$$\frac{\boldsymbol{e}^T \mathbf{Z}_A^T (\boldsymbol{\epsilon} + \beta^0 \boldsymbol{\xi} - \beta^0 \mathbf{Z}_A \boldsymbol{e})}{\sqrt{n}} \xrightarrow{p} 0. \tag{A.16}$$

Then,

$$\frac{\mathbf{X}_0^T (\boldsymbol{\epsilon} + \beta^0 \boldsymbol{\xi} - \beta^0 \mathbf{Z}_A \boldsymbol{e})}{\sqrt{n}} = \frac{\mathbf{X}_0^T (\boldsymbol{\epsilon} + \beta^0 \boldsymbol{\xi})}{\sqrt{n}} - \frac{\mathbf{X}_0^T \beta^0 \mathbf{Z}_A \boldsymbol{e}}{\sqrt{n}},$$

where

$$\begin{aligned} \frac{\mathbf{X}_0^T (\boldsymbol{\epsilon} + \beta^0 \boldsymbol{\xi})}{\sqrt{n}} &\sim N\left(0, (\sigma_1^2 + 2\beta^0 \sigma_{12} + (\beta^0)^2 \sigma_2^2) \frac{\mathbf{X}_0^T \mathbf{X}_0}{n}\right), \\ \frac{\mathbf{X}_0^T \beta^0 \mathbf{Z}_A \boldsymbol{e}}{\sqrt{n}} &\sim N\left(0, (\beta^0)^2 \sigma_2^2 \frac{n}{n_2} \frac{\mathbf{X}_0^T \mathbf{Z}_A}{n} n_2 \mathbf{\Theta} \frac{\mathbf{Z}_A^T \mathbf{X}_0}{n}\right). \end{aligned} \tag{A.17}$$

Note that $\frac{\mathbf{X}_0^T \mathbf{Z}_A}{n} n_2 \mathbf{\Theta} \frac{\mathbf{Z}_A^T \mathbf{X}_0}{n} \to \mathbf{\Psi}_2$. A combination of (A.13), (A.16) and (A.17) yields

that

$$\sqrt{n}\left(\begin{pmatrix}\hat{\beta}^{or}\\ \hat{\boldsymbol{\alpha}}_B^{or}\end{pmatrix}-\begin{pmatrix}\beta^0\\ \boldsymbol{\alpha}_B^0\end{pmatrix}\right)$$
$$\xrightarrow{d} N\left(0,(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2)\boldsymbol{\Sigma}^{-1}+w(\beta^0)^2\sigma_2^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_2\boldsymbol{\Sigma}^{-1}\right).$$

□

Theorem 6. *Under Assumptions 1, 3, 4, 5 and 6, if* $K_2 = p_0$, *then* $P\left((\hat{\beta},\hat{\boldsymbol{\alpha}})=(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or})\right)\to 1$, *either as* $n\to\infty$ *with a fixed* p, *or as both* $n,p\to\infty$.

Proof. By Assumption 1 and 5, it follows from Theorem 5 that the oracle estimator is unique. So

$$\begin{aligned}
I &= P((\hat{\beta},\hat{\boldsymbol{\alpha}})\neq(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or}))\\
&\leq P(\min_{\{\boldsymbol{\alpha}|\frac{1}{\tau}\sum_{j=1}^{p}\min(|\alpha_j|,\tau)\leq K\}\setminus\{\boldsymbol{\alpha}|supp(\boldsymbol{\alpha})=B\}} S(\beta,\boldsymbol{\alpha})\leq S(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or}))\\
&\leq \sum_{B_1\in\mathbb{B}} P(\min_{\{\boldsymbol{\alpha}\left||\boldsymbol{\alpha}_{B_1}|>\tau,|\boldsymbol{\alpha}_{B_1^c}|\leq\tau\right.\}} S(\beta,\boldsymbol{\alpha})\leq S(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or}))\\
&\leq \sum_{B_1\in\mathbb{B}} P(\min_{\{\boldsymbol{\alpha}\left||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\right.\}} S(\beta,\boldsymbol{\alpha})\leq S(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or})).
\end{aligned}$$

For each B_1, we bound $I_{B_1}=P(\min_{\{\boldsymbol{\alpha}\left||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\right.\}} S(\beta,\boldsymbol{\alpha})\leq S(\hat{\beta}^{or},\hat{\boldsymbol{\alpha}}^{or}))$. For $a=n>1$, we have

$$\begin{aligned}
S(\hat{\beta}^T,\hat{\boldsymbol{\alpha}}^T) &\geq \frac{a-1}{a}||\boldsymbol{Y}-\hat{\boldsymbol{D}}\cdot\hat{\beta}^T-\mathbf{Z}_{\mathbf{B_1}}\hat{\boldsymbol{\alpha}}_{B_1}^T||^2-(a-1)||\mathbf{Z}_{B_1^c}\hat{\boldsymbol{\alpha}}_{B_1^c}^T||^2\\
&\geq \frac{a-1}{a}||(\mathbf{I}-\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_{\mathbf{B_1}}})\boldsymbol{Y}||^2-(a-1)\cdot p\cdot c_{max}\tau^2,
\end{aligned}$$

where $(\hat{\beta}^T,\hat{\boldsymbol{\alpha}}^T)=\text{argmin}_{\{\boldsymbol{\alpha}\left||\boldsymbol{\alpha}_{B_1^c}|\leq\tau\right.\}} S(\beta,\boldsymbol{\alpha})$. Rewrite $\boldsymbol{Y}=\beta^0\cdot\hat{\boldsymbol{D}}+\mathbf{Z}_B\boldsymbol{\alpha}_B^0+\boldsymbol{\epsilon}+\beta^0\cdot\boldsymbol{\xi}-\beta^0\cdot\mathbf{Z}_A\boldsymbol{e}$. For simplicity, subsequently, we use notations $\boldsymbol{\epsilon}=\boldsymbol{\epsilon}+\beta^0\cdot\boldsymbol{\xi}-\beta^0\cdot\mathbf{Z}_A\boldsymbol{e}$,

$\boldsymbol{Y} = \mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ and

$$\boldsymbol{\Sigma} = \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2\right)\mathbf{I} + (\beta^0)^2\sigma_2^2 \cdot \mathbf{Z}_A\boldsymbol{\Theta}\mathbf{Z}_A^T = \mathbf{Q}^T\boldsymbol{\Lambda}\mathbf{Q}.$$

Since the eigenvalues of $\frac{\mathbf{Z}_A^T\mathbf{Z}_A}{n}$ are upper-bounded by u_1 and those of $\frac{\mathbf{Z}_A\mathbf{Z}_A^T}{n}$ are also upper-bounded by u_1. Note that the eigenvalues of $\boldsymbol{\Theta}$ are upper-bounded by $\frac{u_2}{n_2}$ and $\frac{n}{n_2}$ is upper-bounded by u_3. So, the eigenvalues of $\boldsymbol{\Lambda}$ are between $\sigma_m^2 = \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2$ and $\sigma_M^2 = \sigma_1^2 + 2\sigma_{12}\beta^0 + \sigma_2^2(\beta^0)^2 + u_1u_2u_3(\beta^0)^2\sigma_2^2$, with $r = \frac{\sigma_M}{\sigma_m} \geq 1$. Now,

$$\begin{aligned}
&S(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) \\
\geq &\frac{a-1}{a}||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\boldsymbol{Y}||^2 - (a-1)\cdot p \cdot c_{max}\tau^2 + \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}})\boldsymbol{Y}.
\end{aligned}$$

Then,

$$S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = ||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B})\boldsymbol{\epsilon}||^2 + \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_B})\boldsymbol{\epsilon}.$$

So,

$$\begin{aligned}
&S(\hat{\beta}^T, \hat{\boldsymbol{\alpha}}^T) - S(\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or}) \\
\geq &-\frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0)^T(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0) \\
&+ (a-1)||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2 - \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B})\boldsymbol{\epsilon} - (a-1)\cdot p \cdot c_{max}\tau^2 \\
&+ \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}})\boldsymbol{Y} - \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup Z_B})\boldsymbol{\epsilon} \equiv -\sum_{j=1}^4 L_j + \sum_{j=1}^4 b_j,
\end{aligned}$$

where $1 > \delta = \delta_1 + \delta_2 + \delta_3$ and $\delta_1, \delta_2, \delta_3 > 0$,

$$L_1 = \frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0)^T(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})(\boldsymbol{\epsilon} - (a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0),$$
$$L_2 = \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B})\boldsymbol{\epsilon},$$
$$L_3 = \frac{a-1}{a}\boldsymbol{Y}^T(\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\boldsymbol{Y},$$
$$L_4 = \boldsymbol{\epsilon}^T(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_B})\boldsymbol{\epsilon},$$
$$b_1 = (a - 1 - \delta)||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2,$$
$$b_2 = \delta_1||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2 - (a-1)\cdot p \cdot c_{max}\tau^2,$$
$$b_3 = \delta_2||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2,$$
$$b_4 = \delta_3||(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2.$$

So, $I_{B_1} \leq P(L_1 \geq b_1) + P(L_2 \geq b_2) + P(L_3 \geq b_3) + P(L_4 \geq b_4)$. For simplicity, we use σ^2 for σ_M^2. Hence, $P(L_1 \geq b_1) \leq \frac{E(e^{\frac{t_1 L_1}{\sigma^2}})}{e^{\frac{t_1 b_1}{\sigma^2}}}$. Let $\boldsymbol{v} = \mathbf{Q}(a-1)(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Z}_B\boldsymbol{\alpha}_B^0$, $\mathbf{P} = \mathbf{Q}(\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}})\mathbf{Q}^T$. Then,

$$\begin{aligned}
&P(L_1 \geq b_1) \\
\leq& e^{-\frac{t_1 b_1}{\sigma^2}} \int \exp\left(\frac{t_1}{a\sigma^2}(\boldsymbol{\epsilon} - \boldsymbol{v})^T(\mathbf{I} - \mathbf{P})(\boldsymbol{\epsilon} - \boldsymbol{v})\right) \cdot (2\pi)^{-n/2}|\boldsymbol{\Lambda}|^{-1/2} \cdot \exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\epsilon}}{2}\right) d\boldsymbol{\epsilon} \\
\leq& e^{-\frac{t_1 b_1}{\sigma^2}} \left(\int \exp\left(\frac{t_1}{a\sigma^2}(\boldsymbol{\epsilon} - \boldsymbol{v})^T(\mathbf{I} - \mathbf{P})(\boldsymbol{\epsilon} - \boldsymbol{v})\right) \cdot (2\pi)^{-n/2}\sigma^{-n} \cdot \exp\left(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2\sigma^2}\right) d\boldsymbol{\epsilon}\right) \cdot r^n \\
\leq& \frac{1}{(1 - 2t_1/a)^{n/2}} \exp\left(-\frac{ni}{d_2\sigma^2}C_{min}\right) \cdot r^n \\
\leq& \frac{1}{(1 - 2t_1/a)^{n/2}} \exp\left(-\frac{ni}{d_2\sigma^2}(C_{min} - d_2\sigma^2 \log r)\right),
\end{aligned}$$

where $i = |B \setminus B_1|$. By Theorem 3 of [2] and Assumption 2, $\sum_{B_1 \in \mathbb{B}} P(L_1 \geq b_1) \to 0$. Similarly, by Assumption 3, $\sum_{B_1 \in \mathbb{B}} P(L_2 \geq b_2) \to 0$.

Now, $\sum_{B_1 \in \mathbb{B}} P(L_3 \geq b_3) \to 0$. Note that $L_3 = 0$ when $A \subseteq B_1$. So,

$$\sum_{B_1 \in \mathbb{B}} P(L_3 \geq b_3) = \sum_{B_1 \in \mathbb{G}} P(L_3 \geq b_3).$$

Hence, $\mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} = \mathbf{P}_{\mathbf{Z}_{\mathbf{B_1}}} + \mathbf{P}_{\boldsymbol{u}}$, here $\boldsymbol{u} = \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{Z}_A \boldsymbol{\gamma}_A^0$; and $\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} = \mathbf{P}_{\mathbf{Z}_{\mathbf{B_1}}} + \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}}$, here $\boldsymbol{v} = \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{Z}_A \boldsymbol{e}$. So, $\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{\mathbf{B_1}}} - \mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} = \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}$. So, for $q = \frac{1}{3}$,

$$\begin{aligned}
&P(L_3 \geq b_3) \\
\leq &P(\boldsymbol{Y}^T (\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}) \boldsymbol{Y} \geq b_3) \\
\leq &P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) + P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \boldsymbol{Y}^T (\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}) \boldsymbol{Y} \geq b_3).
\end{aligned}$$

By Lemma 1, $P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \boldsymbol{Y}^T (\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}) \boldsymbol{Y} \geq b_3) \leq P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \leq \frac{1}{n^q}, \frac{1}{n^q} \boldsymbol{Y}^T \boldsymbol{Y} \geq b_3) \leq P(\frac{1}{n^q} \boldsymbol{Y}^T \boldsymbol{Y} \geq b_3)$. So,

$$P(L_3 \geq b_3) \leq P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) + P(\frac{1}{n^q} \boldsymbol{Y}^T \boldsymbol{Y} \geq b_3).$$

When $B_1 \in \mathbb{B}$,

$$\begin{aligned}
&P(\frac{1}{n^q} \boldsymbol{Y}^T \boldsymbol{Y} \geq b_3) \\
= &P(\frac{||\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n^q} \geq b_3) \\
= &P(\frac{||\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n} \geq \frac{b_3}{n^{1-q}}) \\
\leq &\frac{E(e^{\frac{t}{\sigma^2} \frac{||\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n}})}{e^{\frac{t}{\sigma^2} \frac{b_3}{n^{1-q}}}}
\end{aligned}$$

Then,

$$\begin{aligned}
&E(e^{\frac{t}{\sigma^2} \frac{||\mathbf{Z}_B \boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon}||^2}{n}}) \\
= &\int \exp\left(\frac{t}{\sigma^2} \cdot \frac{(\boldsymbol{\epsilon} + \mathbf{Q}\mathbf{Z}_B \boldsymbol{\alpha}_B^0)^T (\boldsymbol{\epsilon} + \mathbf{Q}\mathbf{Z}_B \boldsymbol{\alpha}_B^0)}{n}\right) (2\pi)^{-n/2} |\boldsymbol{\Lambda}|^{-1/2} \exp\left(-\frac{\boldsymbol{\epsilon}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\epsilon}}{2}\right) d\boldsymbol{\epsilon}
\end{aligned}$$

Let $\boldsymbol{v} = \mathbf{Q}\mathbf{Z}_B\boldsymbol{\alpha}_B^0 = (v_1, \cdots, v_n)$. If $\Lambda_i = \sigma_M^2$. Then,

$$\int \exp\left(\frac{t}{\sigma^2} \cdot \frac{(\epsilon_i + v_i)^2}{n}\right)(2\pi)^{-1/2}\sigma^{-1}\exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \frac{1}{\sqrt{1-2\frac{t}{n}}}\exp\left(\frac{tv_i^2}{\sigma^2(n-2t)}\right).$$

If $\Lambda_i \leq \sigma_M^2$, then

$$\begin{aligned} &\int \exp\left(\tfrac{t}{\sigma^2} \cdot \tfrac{(\epsilon_i + v_i)^2}{n}\right)(2\pi)^{-1/2}\sigma_m^{-1}\exp\left(-\tfrac{\epsilon_i^2}{2\sigma_m^2}\right) \\ =& \tfrac{1}{\sqrt{1-2\frac{t}{nr^2}}}\exp\left(\tfrac{tv_i^2/r^2}{\sigma^2(n-2t/r^2)}\right) \leq \tfrac{1}{\sqrt{1-2\frac{t}{n}}}\exp\left(\tfrac{tv_i^2}{\sigma^2(n-2t)}\right). \end{aligned} \tag{A.18}$$

So,

$$\begin{aligned} &P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3) \leq \frac{E(e^{\frac{t}{\sigma^2}\frac{||\mathbf{Z}_B\boldsymbol{\alpha}_B^0+\boldsymbol{\epsilon}||^2}{n}})}{e^{\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}}} \\ \leq& \frac{(1-2\frac{t}{n})^{-\frac{n}{2}} \cdot e^{\frac{t\cdot||\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{(n-2t)\sigma^2}}}{e^{\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}}} \leq C_1 \cdot e^{-\frac{t}{\sigma^2}\frac{b_3}{n^{1-q}}} \\ =& C_1 \cdot e^{-\frac{t}{\sigma^2}\delta_2 \cdot n^q i C_{\min 1}}, \end{aligned}$$

where C_1 is a positive constant and $i = |B \setminus B_1|$. By similar arguments in Theorem 3 of [2] and by Assumption 2, we have $\sum_{B_1 \in \mathbb{G}} P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3) \leq \sum_{B_1 \in \mathbb{B}} P(\frac{1}{n^q}\boldsymbol{Y}^T\boldsymbol{Y} \geq b_3) \to 0$ as $n, p \to \infty$.

When $B_1 \in \mathbb{G}$,

$$\begin{aligned}
&P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \\
=&P(\boldsymbol{e}^T \mathbf{Z}_A^T \mathbf{M}_{\mathbf{Z}_{\mathbf{B_1}}} \mathbf{Z}_A \boldsymbol{e} \geq \frac{||\boldsymbol{u}||^2}{n^{2q}}) \\
\leq&P(\boldsymbol{e}^T \mathbf{Z}_A^T \mathbf{Z}_A \boldsymbol{e} \geq \frac{||\boldsymbol{u}||^2}{n^{2q}}) \\
\leq&\frac{E(e^{\frac{t}{\sigma_2^2}\boldsymbol{e}^T \mathbf{Z}_A^T \mathbf{Z}_A \boldsymbol{e}})}{e^{\frac{t}{\sigma_2^2}\frac{||u||^2}{n^{2q}}}} \\
\leq&C_2 \cdot e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}},
\end{aligned}$$

where C_2 is a positive constant, $i = |A \setminus B_1|$, and $j = |B_1 \setminus A|$. So,

$$\begin{aligned}
&\sum_{B_1 \in \mathbb{G}} P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \\
\leq&C_2 \sum_{B_1 \in \mathbb{G}} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}} \\
\leq&C_2 \sum_{i=1}^{|A|} \sum_{j=0}^{\max(p_0-|A|+i,0)} \binom{|A|}{i} \binom{p-|A|}{j} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}} \\
\leq&C_2 \sum_{i=1}^{|A|} \sum_{j=0}^{\max(p_0-|A|+i,0)} |A|^i (p-|A|)^j e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}} \\
\leq&C_2 \sum_{i=1}^{|A|} |A|^i \frac{(p-|A|)^{\max(p_0-|A|+i,0)+1} - 1}{p-|A|-1} e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}} \\
\leq&(2C_2 \sum_{i=1}^{|A|} |A|^i (p-|A|)^i e^{-\frac{t}{\sigma_2^2} n^{1-2q} i C_{\min 2}}) \cdot p^{p_0} \\
\leq&2C_2 \frac{|A|(p-|A|) e^{-\frac{t}{\sigma_2^2} n^{1-2q} C_{\min 2}}}{1 - |A|(p-|A|) e^{-\frac{t}{\sigma_2^2} n^{1-2q} C_{\min 2}}} \cdot p^{p_0} \\
\leq&C_3 \cdot \exp\{-\frac{t}{\sigma_2^2} n^{\frac{1}{3}} (C_{\min 2} - \frac{1}{t} \frac{(p_0+2)\log p}{n^{\frac{1}{3}}} \cdot \sigma_2^2)\}.
\end{aligned}$$

By Assumption 4, $\sum_{B_1 \in \mathbb{G}} P(\frac{||\boldsymbol{v}||}{||\boldsymbol{u}||} \geq \frac{1}{n^q}) \to 0$ as $n, p \to \infty$.

To show that $\sum_{B_1 \in \mathbb{B}} P(L_4 \geq b_4) \to 0$, note that $\mathbf{P}_{\mathbf{Z}_A \cup \mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_B} = \mathbf{P}_{\mathbf{U}}$, where $\mathbf{U} = \mathbf{M}_{\mathbf{Z}_B} \mathbf{Z}_A$. Hence,

$$\begin{aligned}
&P(L_4 \geq b_4) \leq P(\boldsymbol{\epsilon}^T \mathbf{P}_{\mathbf{U}} \boldsymbol{\epsilon} \geq b_4) \\
\leq & e^{-\frac{t}{\sigma^2} b_4} \int \exp\left(\frac{t}{\sigma^2} \boldsymbol{\epsilon}^T \mathbf{P}_{\mathbf{U}} \boldsymbol{\epsilon}\right) (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{\boldsymbol{\epsilon}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\epsilon}}{2}\right) d\boldsymbol{\epsilon} \\
\leq & e^{-\frac{t}{\sigma^2} b_4} \left(\int \exp\left(\frac{t}{\sigma^2} \boldsymbol{\epsilon}^T \mathbf{Q} \mathbf{P}_{\mathbf{U}} \mathbf{Q}^T \boldsymbol{\epsilon}\right) (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{2\sigma^2}\right) d\boldsymbol{\epsilon}\right) r^n.
\end{aligned}$$

So,

$$P(L_4 \geq b_4) \leq C_4 \cdot e^{-\frac{t}{\sigma^2} \delta_3 n i (C_{\min 1} - d_2 \sigma^2 \log r)},$$

where C_4 is a positive constant and $i = |B \setminus B_1|$. So, we have $\sum_{B_1 \in \mathbb{B}} P(L_4 \geq b_4) \to 0$, as $n, p \to \infty$. Thus we have proved that $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$ as $n, p \to \infty$. □

Corollary 2. *Assume that Assumptions 1, 3, 4, 5 and 6 are met. If $K_2 = p_0$, then under the null hypothesis of $\beta^0 = 0$, Λ_n converges in distribution to χ^2_1, a chi-squared distribution with degrees of freedom 1, either as $n \to \infty$ with a fixed p, or as both $n, p \to \infty$.*

Proof. By Theorem 6, $P((\hat{\beta}, \hat{\boldsymbol{\alpha}}) \neq (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})) \to 0$. By Theorem 3 of [2], under the null hypothesis, $P(\hat{\boldsymbol{\alpha}}^{(0)}) \neq \hat{\boldsymbol{\alpha}}_0^{or})) \to 0$, where $\hat{\boldsymbol{\alpha}}_0^{or} = \text{argmin}_{\boldsymbol{\alpha}} ||\boldsymbol{Y} - \mathbf{Z}_B \boldsymbol{\alpha}_B||^2$.
As in Theorem 2 of [4], we have

$$\Lambda_n = \frac{n \boldsymbol{\epsilon}^T (\mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_B}) \boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||^2} + \frac{n \boldsymbol{\epsilon}^T (\mathbf{P}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B}) \boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||^2} + R(\boldsymbol{\epsilon}).$$

Hence, $R(\boldsymbol{\epsilon}) \to 0$ as $n \to \infty$. Then, it suffices to show that $\boldsymbol{\epsilon}^T (\mathbf{P}_{\hat{\boldsymbol{D}} \cup Z_B} - \mathbf{P}_{\mathbf{Z}_A \boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_B}) \boldsymbol{\epsilon} \to_p 0$.

Note that $\mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} = \mathbf{P}_{\mathbf{Z}_B} + \mathbf{P}_{\boldsymbol{u}}$ and $\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B} = \mathbf{P}_{\mathbf{Z}_B} + \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}}$, where $\boldsymbol{u} = \mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0$ and $\boldsymbol{v} = \mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{e}$. So, $\mathbf{P}_{\hat{\boldsymbol{D}}\cup\mathbf{Z}_B} - \mathbf{P}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_B} = \mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}}$. Moreover,

$$\begin{aligned}\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{\epsilon} &= \frac{\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v}} - \frac{\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u}} \\ &= \frac{\boldsymbol{\epsilon}^T\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v}} - \frac{(\boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v})\boldsymbol{\epsilon}^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{\epsilon}}{\boldsymbol{u}^T\boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v} + 2\boldsymbol{u}^T\boldsymbol{v})}.\end{aligned}$$

By Assumption 1, $\frac{\boldsymbol{u}^T\boldsymbol{u}}{n} \geq d_0 > 0$. Since $\boldsymbol{e} \sim N(\mathbf{0}, \sigma_2^2\boldsymbol{\Theta})$ and $n_2\boldsymbol{\Theta} \to \boldsymbol{\Theta}_0$, we have $\frac{\boldsymbol{v}^T\boldsymbol{v}}{n} \to_p 0$. Moreover, $\boldsymbol{u}^T\boldsymbol{v}$ follows $N(0, (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\Theta}\mathbf{Z}_A{}^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_2^2)$, so $\frac{\boldsymbol{u}^T\boldsymbol{v}}{n} \to_p 0$. Again, $\boldsymbol{\epsilon}^T\boldsymbol{u}$ follows $N(0, (\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_1^2)$. So $\frac{\boldsymbol{\epsilon}^T\boldsymbol{u}}{\sqrt{n}}$ has the variance no greater than $\frac{(\boldsymbol{\gamma}_A^0)^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{\gamma}_A^0\sigma_1^2}{n}$. And $\boldsymbol{\epsilon}^T\boldsymbol{v} = \boldsymbol{\epsilon}^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A\boldsymbol{e}$, $\boldsymbol{e} \to_p 0$. Finally, $\boldsymbol{\epsilon}^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A$ follows $N(0, \mathbf{Z}_A^T\mathbf{M}_{\mathbf{Z}_B}\mathbf{Z}_A)$. So, $\frac{\boldsymbol{\epsilon}^T\boldsymbol{v}}{\sqrt{n}} \to_p 0$. Consequently, $\boldsymbol{\epsilon}^T(\mathbf{P}_{\boldsymbol{u}+\boldsymbol{v}} - \mathbf{P}_{\boldsymbol{u}})\boldsymbol{\epsilon} \to_p 0$ as $n \to \infty$. Thus $\Lambda_n \to_d \chi_1^2$. □

Theorem 7. *Assume that Assumptions 1, 3, 4, 5 and 6 are met. Then when p is fixed, if $p_0 \in \mathcal{K}_2$, we have $P\left(\hat{K}_2 = p_0\right) \to 1$ as $n \to \infty$.*

Proof. It is sufficient to show $P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_0)\right) \to 0$ for any $K_2 \neq p_0$. From Theorem 6, we have $P(\hat{\boldsymbol{\alpha}}_{p_0} = \hat{\boldsymbol{\alpha}}^{or}) \to 1$, so $P(||\hat{\boldsymbol{\alpha}}_{p_0}||_0 = p_0) \to 1$. We have

$$\begin{aligned}\mathrm{BIC}(p_0) &= n \cdot \log\frac{||\boldsymbol{Y} - \hat{\beta}_{p_0}\cdot\hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{p_0}||^2}{n} + \log(n)\cdot||\hat{\boldsymbol{\alpha}}_{p_0}||_0 \\ &\leq n \cdot \log\frac{||\boldsymbol{Y} - \beta^0\cdot\hat{\boldsymbol{D}} - \mathbf{Z}\boldsymbol{\alpha}^0||^2}{n} + \log(n)\cdot p_0 \\ &= n \cdot \log\frac{||\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi} + \beta^0\mathbf{Z}_A\boldsymbol{e}||^2}{n} + \log(n)\cdot p_0,\end{aligned}$$

here $\boldsymbol{e} = \boldsymbol{\gamma}_A^0 - \hat{\boldsymbol{\gamma}}_A \sim N(\mathbf{0}, \sigma_2^2\boldsymbol{\Theta})$, and $||\boldsymbol{e}||^2 = o_p(1)$ by Assumption 5. Suppose for some $B_1 \subseteq S$, $|B_1| = K_2$ and $|\,(\hat{\boldsymbol{\alpha}}_{K_2})_{B_1^c}\,| \leq \tau_2$, with Assumption 3 and similar arguments

in proof of Theorem 6, we have

$$\begin{aligned}\mathrm{BIC}(K_2) &= n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{K_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}_{K_2}||^2}{n} + \log(n) \cdot ||\hat{\boldsymbol{\alpha}}_{K_2}||_0 \\ &\geq n \cdot \log \frac{||\boldsymbol{Y} - \hat{\beta}_{K_2} \cdot \hat{\boldsymbol{D}} - \mathbf{Z}_{B_1}(\hat{\boldsymbol{\alpha}}_{K_2})_{B_1}||^2 - n \cdot p \cdot c_{max}\tau_2^2}{n} + \log(n) \cdot K_2 \\ &\geq n \cdot \log \frac{||\mathbf{M}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2.\end{aligned}$$

So we have

$$\begin{aligned}&P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_0)\right) \\ &\leq P\left(n \cdot \log \frac{||\mathbf{M}_{\hat{\boldsymbol{D}} \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \leq n \cdot \log \frac{||\boldsymbol{\epsilon} + \beta^0 \boldsymbol{\xi} + \beta^0 \mathbf{Z}_A \boldsymbol{e}||^2}{n} + \log(n) \cdot p_0\right).\end{aligned}$$

Since $||\boldsymbol{e}||^2 = o_p(1)$, and $\hat{\boldsymbol{D}} = \mathbf{Z}_A\boldsymbol{\gamma}_A^0 - \mathbf{Z}_A\boldsymbol{e}$, with Assumption 4 and similar arguments in proof of Theorem 6, we get

$$\begin{aligned}&P(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_0)) \\ &\leq P(n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \boldsymbol{Y}||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \leq n \cdot \log \frac{||\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}||^2}{n} + \log(n) \cdot p_0) \\ &= P(n \cdot \log \frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} (\mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})||^2 - 6\sigma_M^2}{n} + \log(n) \cdot K_2 \\ &\leq n \cdot \log \frac{||\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi}||^2}{n} + \log(n) \cdot p_0).\end{aligned}$$

We have

$$\begin{aligned}&\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} (\mathbf{Z}_B\boldsymbol{\alpha}_B^0 + \boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})||^2 - 6\sigma_M^2}{n} \\ =&\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n} + \frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} (\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})||^2}{n} \\ &+ 2 \cdot \frac{(\boldsymbol{\epsilon} + \beta^0\boldsymbol{\xi})^T \mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0 \cup \mathbf{Z}_{B_1}} \mathbf{Z}_B\boldsymbol{\alpha}_B^0 - 6\sigma_M^2}{n},\end{aligned}$$

and the last term

$$\frac{\left(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}\right)^T\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0-6\sigma_M^2}{n}$$
$$\sim N\left(\frac{-6\sigma_M^2}{n},\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\frac{(\boldsymbol{\alpha}_B^0)^T\mathbf{Z}_B^T\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0}{n^2}\right)=o_p(1),$$

so we have

$$\begin{aligned}&P\left(\mathrm{BIC}(K_2)\le\mathrm{BIC}(p_0)\right)\\ \le\ &P(n\log(\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n}+\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi})||^2}{n}+o_p(1))+\log(n)K_2\\ &\le n\log\frac{||\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}||^2}{n}+\log(n)p_0),\end{aligned}$$

here

$$\begin{aligned}\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\left(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}\right)||^2}{n}&\sim\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\frac{\chi^2_{n-K_2-1}}{n}\\ &\to\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right),\end{aligned}$$

$$\frac{||\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}||^2}{n}\sim\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right)\frac{\chi^2_n}{n}\to\left(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2\right).$$

When $K_2<p_0$, with Assumption 6, we have

$$\liminf_{n\to\infty}\frac{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}\mathbf{Z}_B\boldsymbol{\alpha}_B^0||^2}{n}=C,$$

for some positive constant C, thus

$$\begin{aligned}&P\left(\mathrm{BIC}(K_2)\le\mathrm{BIC}(p_0)\right)\\ =&P\left(n\cdot\log(1+\frac{n\cdot C}{(\sigma_1^2+2\sigma_{12}\beta^0+(\beta^0)^2\sigma_2^2)\,\chi_n^2}+o_p(1))\le\log(n)\cdot(p_0-K_2)\right)\to 0.\end{aligned}\tag{A.19}$$

When $K_2 > p_0$, we have

$$\begin{aligned} & P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_0)\right) \\ \leq & P\left(\log(n)\cdot(K_2-p_0) \leq n\cdot\log\frac{||\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi}||^2}{||\mathbf{M}_{\mathbf{Z}_A\boldsymbol{\gamma}_A^0\cup\mathbf{Z}_{B_1}}(\boldsymbol{\epsilon}+\beta^0\boldsymbol{\xi})||^2-6\sigma_M^2}\right) \\ = & P\left(\log(n)\cdot(K_2-p_0) \leq n\cdot\log(1+\frac{\chi^2_{K_2+1}+6\sigma_M^2}{\chi^2_{n-K_2-1}-6\sigma_M^2})\right) \to 0. \end{aligned} \tag{A.20}$$

Combining (A.19) and (A.20) we get $P\left(\mathrm{BIC}(K_2) \leq \mathrm{BIC}(p_0)\right) \to 0$ for any $K_2 \neq p_0$, thus $P\left(\hat{K}_2 = p_0\right) \to 1$ as $n \to \infty$. □

Appendix B

Consistent Estimation of the Asymptotic Variances

B.1 One-Sample Case

B.1.1 Estimating v with the Oracle Estimators

For the purpose of presentation, we first consider estimating the asymptotic variance v in Theorem 1 using the Oracle estimators. Note that

$$v = \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right) \cdot \boldsymbol{\Sigma}_{11}^{-1} - \left(2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right) \cdot \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)_{11}.$$

Since $\hat{\boldsymbol{\gamma}}_A^{or} = \boldsymbol{\gamma}_A^0 + (\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi}$ and $(\mathbf{Z}_A^T\mathbf{Z}_A)^{-1}\mathbf{Z}_A^T\boldsymbol{\xi} \xrightarrow{p} \mathbf{0}_{|A|\times 1}$, we have $\hat{\boldsymbol{\gamma}}_A^{or} \xrightarrow{p} \boldsymbol{\gamma}_A^0$ as well as

$$\frac{\mathbf{Z}_A^T\mathbf{Z}_A}{n} \to \mathbf{U}_A, \quad \frac{\mathbf{Z}_A^T\mathbf{Z}_B}{n} \to \mathbf{U}_{AB}, \quad \frac{\mathbf{Z}_B^T\mathbf{Z}_B}{n} \to \mathbf{U}_B,$$

as $n \to \infty$. Plug in these consistent estimates, we obtain consistent estimators of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ as

$$\begin{aligned}\hat{\mathbf{\Sigma}}^{or} &= \begin{pmatrix} (\hat{\boldsymbol{\gamma}}_A^{or})^T \mathbf{Z}_A^T \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A^{or} & (\hat{\boldsymbol{\gamma}}_A^{or})^T \mathbf{Z}_A^T \mathbf{Z}_B \\ \mathbf{Z}_B^T \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A^{or} & \mathbf{Z}_B^T \mathbf{Z}_B \end{pmatrix} /n \xrightarrow{p} \mathbf{\Sigma}, \\ \hat{\mathbf{\Psi}}^{or} &= \begin{pmatrix} (\hat{\boldsymbol{\gamma}}_A^{or})^T \mathbf{Z}_A^T \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A^{or} & (\hat{\boldsymbol{\gamma}}_A^{or})^T \mathbf{Z}_A^T \mathbf{Z}_B \\ \mathbf{Z}_B^T \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A^{or} & \mathbf{Z}_B^T \mathbf{P}_{\mathbf{Z}_A} \mathbf{Z}_B \end{pmatrix} /n \xrightarrow{p} \mathbf{\Psi}. \end{aligned} \tag{B.1}$$

To estimate $(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2)$, let $\hat{\boldsymbol{D}}^{or} = \mathbf{Z}_A \hat{\boldsymbol{\gamma}}_A^{or}$ and $\mathbf{X} = (\hat{\boldsymbol{D}}^{or}, \mathbf{Z}_B)$. Then,

$$\hat{\boldsymbol{Y}}^{or} = \hat{\beta}^{or} \cdot \hat{\boldsymbol{D}}^{or} + \mathbf{Z}_B \hat{\boldsymbol{\alpha}}_B^{or} = \mathbf{P}_{\mathbf{X}} \boldsymbol{Y}.$$

Let

$$\hat{v}_1^{or} = \frac{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{or}||^2}{n} = \frac{\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)^T (\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)}{n}.$$

From (A.7), as $n \to \infty$,

$$\frac{\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)^T \left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)}{n} \xrightarrow{p} \sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2.$$

By $\mathbf{X}^T\mathbf{X}/n \xrightarrow{p} \mathbf{\Sigma}$ and (A.1), we have

$$\begin{aligned} &\frac{\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)^T \mathbf{P}_{\mathbf{X}} \left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)}{n} \\ =&\frac{\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)^T \mathbf{X}}{n} \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)^{-1} \frac{\mathbf{X}^T \left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right)}{n} \xrightarrow{p} 0. \end{aligned}$$

So,

$$\hat{v}_1^{or} \xrightarrow{p} \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right). \tag{B.2}$$

Next, we estimate σ_1^2. Let $\mathbf{X}^* = (\boldsymbol{D}, \mathbf{Z}_B)$ and

$$\hat{\boldsymbol{Y}}_*^{or} = \hat{\beta}^{or} \cdot \boldsymbol{D} + \mathbf{Z}_B \hat{\boldsymbol{\alpha}}_B^{or},$$

we have

$$\boldsymbol{Y} - \hat{\boldsymbol{Y}}_*^{or} = \boldsymbol{\epsilon} - \mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\beta^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_A})\boldsymbol{\xi} + \boldsymbol{\epsilon}\right).$$

Define

$$\hat{v}_2^{or} = \frac{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_*^{or}||^2}{n}.$$

Note that $\mathbf{X}^T\mathbf{X}/n \xrightarrow{p} \boldsymbol{\Sigma}$, $\boldsymbol{\epsilon}^T\mathbf{X}^*/n \xrightarrow{p} 0$, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n \xrightarrow{p} \sigma_1^2$. By (A.1), we have

$$\hat{v}_2^{or} \xrightarrow{p} \sigma_1^2. \tag{B.3}$$

Combining (B.1), (B.2) and (B.3) yields a consistent estimator $\hat{v}^{or}$ of v in Theorem 1 as follows:

$$\hat{v}^{or} = \hat{v}_1^{or} \cdot (\hat{\boldsymbol{\Sigma}}^{or})_{11}^{-1} - (\hat{v}_1^{or} - \hat{v}_2^{or}) \cdot \left((\hat{\boldsymbol{\Sigma}}^{or})^{-1}\hat{\boldsymbol{\Psi}}^{or}(\hat{\boldsymbol{\Sigma}}^{or})^{-1}\right)_{11} \xrightarrow{p} v.$$

B.1.2 Estimating v with the 2ScML Estimators

In practice, since the Oracle estimators are unknown, we estimate v using the 2ScML estimators $\hat{\boldsymbol{\gamma}}$, $\hat{\beta}$ and $\hat{\boldsymbol{\alpha}}$. Let $\hat{A} = \{i \big| \hat{\gamma}_i \neq 0, 1 \leq i \leq p\}$ and $\hat{B} = \{j \big| \hat{\alpha}_j \neq 0, 1 \leq j \leq p\}$. In parallel with $\hat{\boldsymbol{\Sigma}}^{or}$ and $\hat{\boldsymbol{\Psi}}^{or}$, we use

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \begin{pmatrix} \hat{\boldsymbol{\gamma}}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}} & \hat{\boldsymbol{\gamma}}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}^T\mathbf{Z}_{\hat{B}} \\ \mathbf{Z}_{\hat{B}}^T\mathbf{Z}_{\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}} & \mathbf{Z}_{\hat{B}}^T\mathbf{Z}_{\hat{B}} \end{pmatrix} /n, \\
\hat{\boldsymbol{\Psi}} &= \begin{pmatrix} \hat{\boldsymbol{\gamma}}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}} & \hat{\boldsymbol{\gamma}}_{\hat{A}}^T\mathbf{Z}_{\hat{A}}^T\mathbf{Z}_{\hat{B}} \\ \mathbf{Z}_{\hat{B}}^T\mathbf{Z}_{\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}} & \mathbf{Z}_{\hat{B}}^T\mathbf{P}_{\mathbf{Z}_{\hat{A}}}\mathbf{Z}_{\hat{B}} \end{pmatrix} /n.
\end{aligned} \tag{B.4}$$

Let $\hat{\boldsymbol{D}} = \mathbf{Z}_{\hat{A}}\hat{\boldsymbol{\gamma}}_{\hat{A}}$. Then

$$\hat{\boldsymbol{Y}} = \hat{\beta} \cdot \hat{\boldsymbol{D}} + \mathbf{Z}_{\hat{B}}\hat{\boldsymbol{\alpha}}_{\hat{B}}.$$

Now, we replace $\hat{v}_1^{or}$ by

$$\hat{v}_1 = \frac{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2}{n}. \tag{B.5}$$

Let

$$\hat{\boldsymbol{Y}}_* = \hat{\beta} \cdot \boldsymbol{D} + \mathbf{Z}_{\hat{B}}\hat{\boldsymbol{\alpha}}_{\hat{B}}.$$

Then, we replace $\hat{v}_2^{or}$ by

$$\hat{v}_2 = \frac{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_*||^2}{n}. \tag{B.6}$$

A combination of (B.4), (B.5) and (B.6) yields that

$$\hat{v} = \hat{v}_1 \cdot (\hat{\boldsymbol{\Sigma}})_{11}^{-1} - (\hat{v}_1 - \hat{v}_2) \cdot \left(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Sigma}}^{-1}\right)_{11}.$$

By Theorem 2, $P\left(\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}^{or}\right) \to 1$, $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$, $P(\hat{A} = A) \to 1$, and $P(\hat{B} = B) \to 1$. Thus, $P(\hat{v} = \hat{v}^{or}) \to 1$, implying that $\hat{v} \xrightarrow{p} v$.

B.2 Two-Sample Case

B.2.1 Estimating v with the Oracle Estimators

By Theorem 4, the asymptotic variance v is

$$v = \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right) \cdot \boldsymbol{\Sigma}_{11}^{-1} + w(\beta^0)^2\sigma_2^2(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_2\boldsymbol{\Sigma}^{-1})_{11}.$$

As in the one-sample case, let $\hat{\boldsymbol{D}} = \mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A$ and $\mathbf{X} = (\hat{\boldsymbol{D}}, \mathbf{Z}_B)$. Then,

$$\hat{\boldsymbol{Y}}^{or} = \hat{\beta}^{or} \cdot \hat{\boldsymbol{D}} + \mathbf{Z}_B\hat{\boldsymbol{\alpha}}_B^{or} = \mathbf{P}_{\mathbf{X}}\boldsymbol{Y},$$
$$\hat{v}_1^{or} = \frac{||\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{or}||^2}{n} \xrightarrow{p} \left(\sigma_1^2 + 2\sigma_{12}\beta^0 + (\beta^0)^2\sigma_2^2\right).$$

Given the second sample of size n_2, some consistent estimators $\hat{\boldsymbol{\Theta}}_0$ and $\hat{\sigma}_2^2$ of $\boldsymbol{\Theta}_0$ and σ_2^2 can be obtained, for example, by linear regression. As in the one-sample case, $\hat{\boldsymbol{\Sigma}}^{or}$ and $\hat{\boldsymbol{\Psi}}_2^{or}$ are consistent estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}_2$, respectively, that is,

$$\hat{\boldsymbol{\Sigma}}^{or} = \begin{pmatrix} \hat{\boldsymbol{\gamma}}_A^T\mathbf{Z}_A^T\mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A & \hat{\boldsymbol{\gamma}}_A^T\mathbf{Z}_A^T\mathbf{Z}_B \\ \mathbf{Z}_B^T\mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A & \mathbf{Z}_B^T\mathbf{Z}_B \end{pmatrix} /n \xrightarrow{p} \boldsymbol{\Sigma},$$
$$\hat{\boldsymbol{\Psi}}_2^{or} = \begin{pmatrix} \hat{\boldsymbol{\gamma}}_A^T\mathbf{Z}_A^T\mathbf{Z}_A/n \\ \mathbf{Z}_B^T\mathbf{Z}_A/n \end{pmatrix} \hat{\boldsymbol{\Theta}}_0 \begin{pmatrix} \mathbf{Z}_A^T\mathbf{Z}_A\hat{\boldsymbol{\gamma}}_A/n & \mathbf{Z}_A^T\mathbf{Z}_B/n \end{pmatrix} \xrightarrow{p} \boldsymbol{\Psi}_2.$$

In (Appendix B.2.1), we estimate w by n/n_2 and β_0 by $\hat{\beta}^{or}$. Finally, $\hat{v}^{or}$ of v can be written as

$$\hat{v}^{or} = \hat{v}_1^{or} \cdot (\hat{\boldsymbol{\Sigma}}^{or})_{11}^{-1} + \frac{n}{n_2}(\hat{\beta}^{or})^2\hat{\sigma}_2^2 \left((\hat{\boldsymbol{\Sigma}}^{or})^{-1}\hat{\boldsymbol{\Psi}}_2^{or}(\hat{\boldsymbol{\Sigma}}^{or})^{-1}\right)_{11} \xrightarrow{p} v,$$

as $n \to \infty$.

B.2.2 Estimating v with the 2ScML estimators

Since the Oracle estimators are unknown in practice, we estimate v with the 2ScML estimators $\hat{\beta}$ and $\hat{\boldsymbol{\alpha}}$. Similar to the one-sample case, in $\hat{v}^{or}$, we replace $\hat{\beta}^{or}$, $\hat{\boldsymbol{\alpha}}^{or}$ and B with $\hat{\beta}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{B}$ respectively, to obtain $\hat{v}$. By Theorem 5, we have $P\left((\hat{\beta}, \hat{\boldsymbol{\alpha}}) = (\hat{\beta}^{or}, \hat{\boldsymbol{\alpha}}^{or})\right) \to 1$, and $P(\hat{B} = B) \to 1$, implying $P(\hat{v} = \hat{v}^{or}) \to 1$, and thus a consistent estimator $\hat{v} \xrightarrow{p} v$.

Appendix C

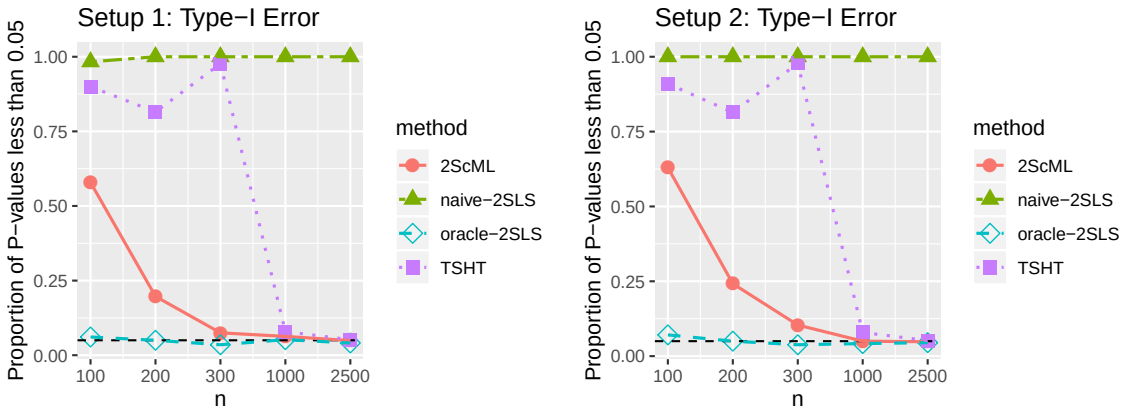
More Simulation Results

C.1 Full Simulation 1 Results: One-Sample Case with IV Assumptions (A) and (C) Violated

We compared 2ScML with TSHT, naive-2SLS, and oracle-2SLS through simulations, and the simulation setups closely followed those of TSHT [1]. First we compared the Type-I Errors of these methods. In Setup 1, we set the number of IVs $p = 100$, and the sample size n varying at $100, 200, 300, 1000$ and 2500. Instruments $\boldsymbol{Z}$'s followed a multivariate normal distribution with mean 0 and an AR(0.5) covariance matrix $\boldsymbol{\Sigma}$: $\boldsymbol{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. The error terms (ϵ, ξ) were generated from a bivariate normal distribution with mean 0, variance 1.5 and covariance 0.75. For $2 \leq i \leq 8$, $\gamma_i^0 = 0.5$; otherwise $\gamma_i^0 = 0$; i.e. the 2^{nd} to 8^{th} IVs were relevant and had an equal effect size 0.5. For $i = 7, 8$, $\alpha_i^0 = 0.5$; otherwise $\alpha_i^0 = 0$; i.e. the 7^{th} and 8^{th} instruments were invalid IVs and had some direct effects on the outcome. When $\beta^0 = 0$, there was no causal effect from the exposure to outcome, i.e. the null case. Setup 2 was the same as Setup 1 except: for $i = 1, 7, 8, 9$, $\alpha_i^0 = 0.5$; otherwise $\alpha_i^0 = 0$; i.e. the relevant 7^{th} and 8^{th} instruments were invalid, and the irrelevant 1^{st} and 9^{th} instruments were also invalid.

For both setups, in each simulation we generated n samples from the model in (1),

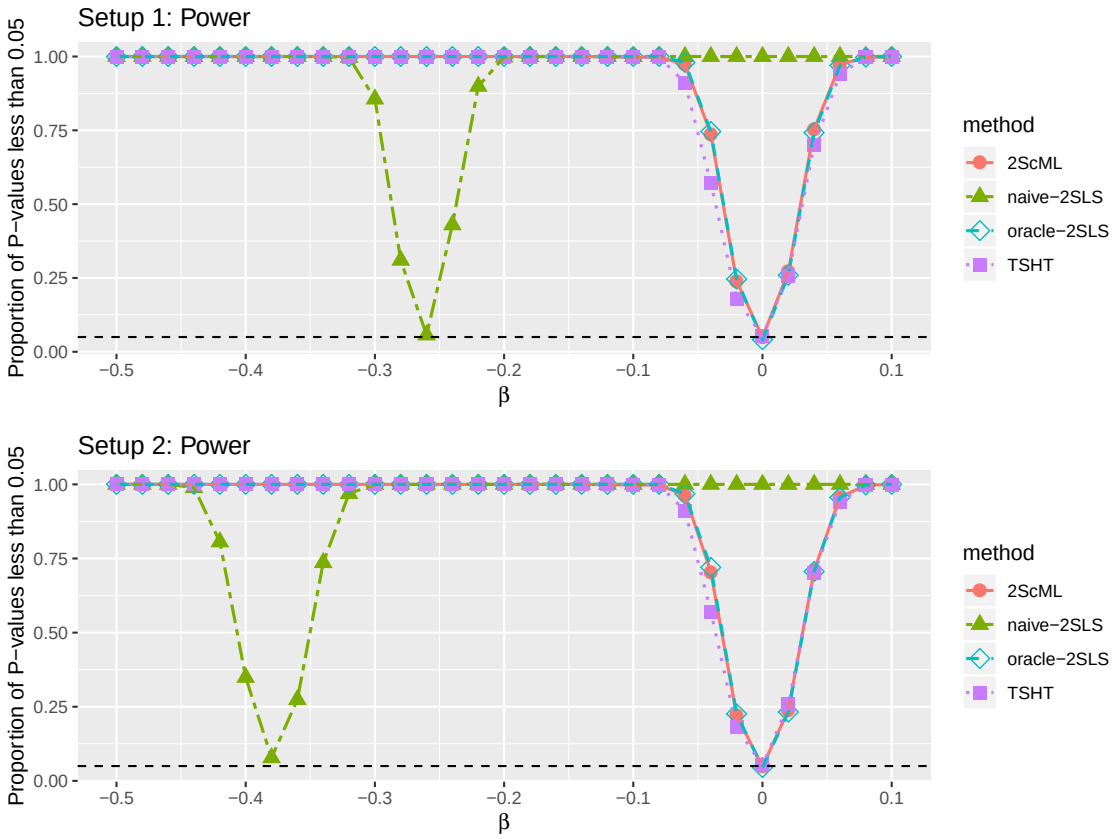
Figure C.1: Empirical Type-I Error Rates of Setup 1 and Setup 2: the x-axis shows the sample size, while y-axis shows the empirical Type-I Error rates based on 1000 simulations; the horizontal dashed line represents the nominal level 0.05.



then we applied the four methods to simulated data to test $H_0 : \beta^0 = 0$ versus $H_1 : \beta^0 \neq 0$. For 2ScML, in stage 1 we used BIC to choose the best $K_1 \in \{5, 6, 7, 8, 9, 10\}$; in stage 2 we used BIC to choose the best $K_2 \in \{0, 1, 2, 3, 4, 5\}$; and we set $\tau_1 = \tau_2 = 1 \times 10^{-5}$. For the naive-2SLS, in stage 1 we used the 2^{nd} to 8^{th} IVs to get $\hat{\boldsymbol{D}}$, then in stage 2 we fitted a linear regression model of Y on $\hat{\boldsymbol{D}}$. The oracle-2SLS had the same stage 1 as the naive-2SLS, but in stage 2 we fitted a linear regression model of Y on $\hat{\boldsymbol{D}}$ and the 7^{th} and 8^{th} IVs for Setup 1, or $\hat{\boldsymbol{D}}$ and the 1^{st}, 7^{th}, 8^{th} and 9^{th} IVs for Setup 2; in other words, we included the 2 and 4 invalid IVs for Setup 1 and Setup 2 respectively.

For each setup, we repeated the simulation 1000 times and set the nominal significant level at 0.05 for each n; Figure C.1 shows the simulation result. We can see that the oracle-2SLS could always have a Type-I Error rate around the nominal level 0.05, while naive-2SLS had a Type-I Error rate dramatically inflated around 1. When the sample size was small, TSHT and 2ScML both had large Type-I Error rates. But as the sample size increased from 100 to 300, the Type-I Error rate of 2ScML decreased fast, while that of TSHT still had a relatively large Type-I Error. When the sample

Figure C.2: Empirical Power Rates of Setup 1 and Setup 2: the x-axis shows the causal effect size β^0, while y-axis shows the empirical power rate based on 1000 simulations; the horizontal dashed line represents the nominal level 0.05.



size was large enough, both TSHT and 2ScML could control their Type-I Error rate satisfactorily around 0.05.

From Figure C.1 we can see that when the sample size was 2500, 2ScML, TSHT and oracle-2SLS could control their Type-I Error rates. So we compared their power at the sample size 2500, and in both setups we changed β^0 from -0.5 to 0.1 with a step size of 0.02. Then we applied all 4 methods and repeated the simulation 1000 times to calculate their empirical power. Figure C.2 shows the results with the x-axis representing value of β^0. Again, when $\beta^0 = 0$, i.e. with no causal effect, 2ScML, TSHT and oracle-2SLS could control Type-I Error at 0.05, while naive-2SLS had a

Table C.1: Empirical type I error rates (for $\beta^0 = 0$) and power (for $\beta^0 \neq 0$) of the methods in Simulation 1.

β^0	Setup 1			Setup 2		
	oracle-2SLS	2ScML	TSHT	oracle-2SLS	2ScML	TSHT
-0.10	1.000	0.998	0.999	1.000	0.998	0.999
-0.08	1.000	0.998	0.998	1.000	0.998	0.998
-0.06	0.980	0.973	0.911	0.968	0.963	0.911
-0.04	0.746	0.738	0.572	0.720	0.705	0.571
-0.02	0.246	0.239	0.180	0.226	0.221	0.180
0.00	0.041	0.049	0.052	0.045	0.048	0.052
0.02	0.259	0.271	0.258	0.232	0.239	0.258
0.04	0.742	0.754	0.701	0.706	0.710	0.700
0.06	0.970	0.973	0.941	0.956	0.957	0.941
0.08	0.998	0.997	0.997	0.998	0.997	0.997
0.10	1.000	0.999	0.999	1.000	1.000	0.999

Type-I Error inflated to 1. In both Setups 1 and 2, when $|\beta^0|$ was large, 2ScML, TSHT and oracle-2SLS all had power 1. When $|\beta^0|$ was small, the power of 2ScML and that of oracle-2SLS were very close, and typically higher than that of TSHT. Table C.1 shows the power of 2ScML, TSHT and oracle-2SLS for $-0.1 \leq \beta^0 \leq 0.1$. This supports the theory that 2ScML has the oracle property while TSHT does not. We can see that, for the range $-0.3 < \beta^0 < -0.22$ in Setup 1, naive-2SLS had power smaller than 1, while the other three methods had power 1. This was not surprising: the direct effects $\boldsymbol{\gamma}^0$ and $\boldsymbol{\alpha}^0$ were positive, and the total effect of IVs on Y was $(\beta^0 \cdot \boldsymbol{\gamma}^0 + \boldsymbol{\alpha}^0)$, so negative β^0 in a certain range would diminish the total effect toward 0 and decrease the power of naive-2SLS. We can see the similar pattern in Setup 2, while the range was $-0.42 < \beta^0 < -0.32$ with larger absolute values as compared to $-0.3 < \beta^0 < -0.22$ in Setup 1. The reason was that we had two more invalid IVs, the 1^{st} and 9^{th} SNPs, with positive direct effects on Y, so it took negative β^0 with larger absolute values to cancel out the total effect.

C.2 Simulation 3: Setups Mimicking Real Data

We did simulations with some realistic setups to mimic real data applications. We first performed 2ScML to study the causal effect of gene GEMIN7 on chromosome 19 on Alzheimer's disease (AD) and generated simulated data using the estimated parameters and real SNP data. More specifically, from the TWAS Fusion database, 366 SNPs from position 45092942 to 46094597 on chromosome 19 were used to estimate the expression level of GEMIN7; then from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, we obtained the individual-level genotype data, i.e. values of SNPs, and the outcome indicating the AD status for 785 individuals. We extracted all SNPs from 44992000 to 46195000 on chromosome 19 from the ADNI data, i.e. by roughly extending 100kb up- and down-stream the gene region. After cleaning and pruning out SNPs to make their pairwise correlations less than 0.6, we had 253 SNPs left. Among these 253 SNPs, 42 overlapped with the 366 SNPs with non-zero estimated coefficients in the TWAS Fusion database, and 10 of them had the absolute values of their estimated coefficients larger than 0.003 as shown in Table C.2. We used these 42 SNPs with their TWAS Fusion estimated coefficients to predict gene-expression levels in the first stage. In the second stage, we applied 2ScML: we included the predicted gene-expression and all 253 SNPs; using cross-validation we chose the best K_2 as 3. The 3 chosen SNPs with their estimated coefficients were: rs8100875 (position 45007076, coefficient -0.16), noRSname (position 45386467, coefficient 0.12), rs2288918 (position 45528799, coefficient 0.10). We show the correlations between these 3 SNPs and the 10 SNPs relevant to the expression level of GEMIN7 in Table C.2.

We then generated the simulated data accordingly. We used the real genotypes of the 253 SNPs from the 785 individuals in ADNI, so we had the number of instruments as $p = 253$ and the sample size $n = 785$ for each simulated dataset. From the

Table C.2: Ten relevant IVs/SNPs on chromosome 19 used to generate exposure D, and their correlations with invalid IVs/SNPs rs8100875, noRSname and rs2288918.

RS Name	Position	γ^0	Correlation with rs8100875	Correlation with noRSname	Correlation with rs2288918
rs17658470	45115393	-0.00356	-0.0541	-0.0116	0.00677
rs2965164	45202052	-0.00471	-0.0183	-0.127	-0.0594
rs10421830	45210634	-0.00393	0.0639	-0.099	-0.00137
rs10405693	45326664	-0.00434	-0.0297	0.202	0.0431
rs440277	45361224	0.00972	0.0578	-0.000728	-0.0641
rs283814	45389224	0.00319	-0.00327	-0.147	-0.0438
rs10405859	45602781	0.0184	0.0204	-0.135	-0.283
rs238419	45853413	0.0103	0.0280	0.0241	-0.0150
rs8099878	46019601	-0.0119	-0.0194	0.0487	0.0801
rs8111589	46034558	-0.00384	0.0495	0.0398	0.138

TWAS Fusion website, the proportion of the variance of the expression level for gene GEMIN7 explained by SNPs was about 10%. We had $var(-0.0035 \times rs17658470 + \cdots + -0.0038 \times rs8111589) = 0.000436$, so we generated ξ from a normal distribution with mean 0 and variance 0.004. In stage 1 we generated exposure D with model of (1) using 10 SNPs with their coefficients shown in Table C.2 and ξ. The proportion of variance of outcome Y explained by 3 SNPs was around 10%, and $var(-0.16 \times rs8100875 + 0.12 \times noRSname + 0.10 \times rs2288918) = 0.014$, so we generated ϵ from a normal distribution with mean 0 and variance 0.12, and set its correlation with ξ as 0.2. For Setup 1, in stage 2 we generated Y from the model in (1) using generated D, the 3 invalid IVs/SNPs, and ϵ. The direct effects α's of these 3 SNPs were -0.16, 0.12, 0.10 respectively, and the causal effect β of D was one of values at -2, -1, -0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 1 and 2. For Setup 2, in stage 2 we excluded the 3 invalid IVs/SNPs while other aspects remained the same as in Setup 1, so we had 10 SNPs relevant to D and all SNPs were valid.

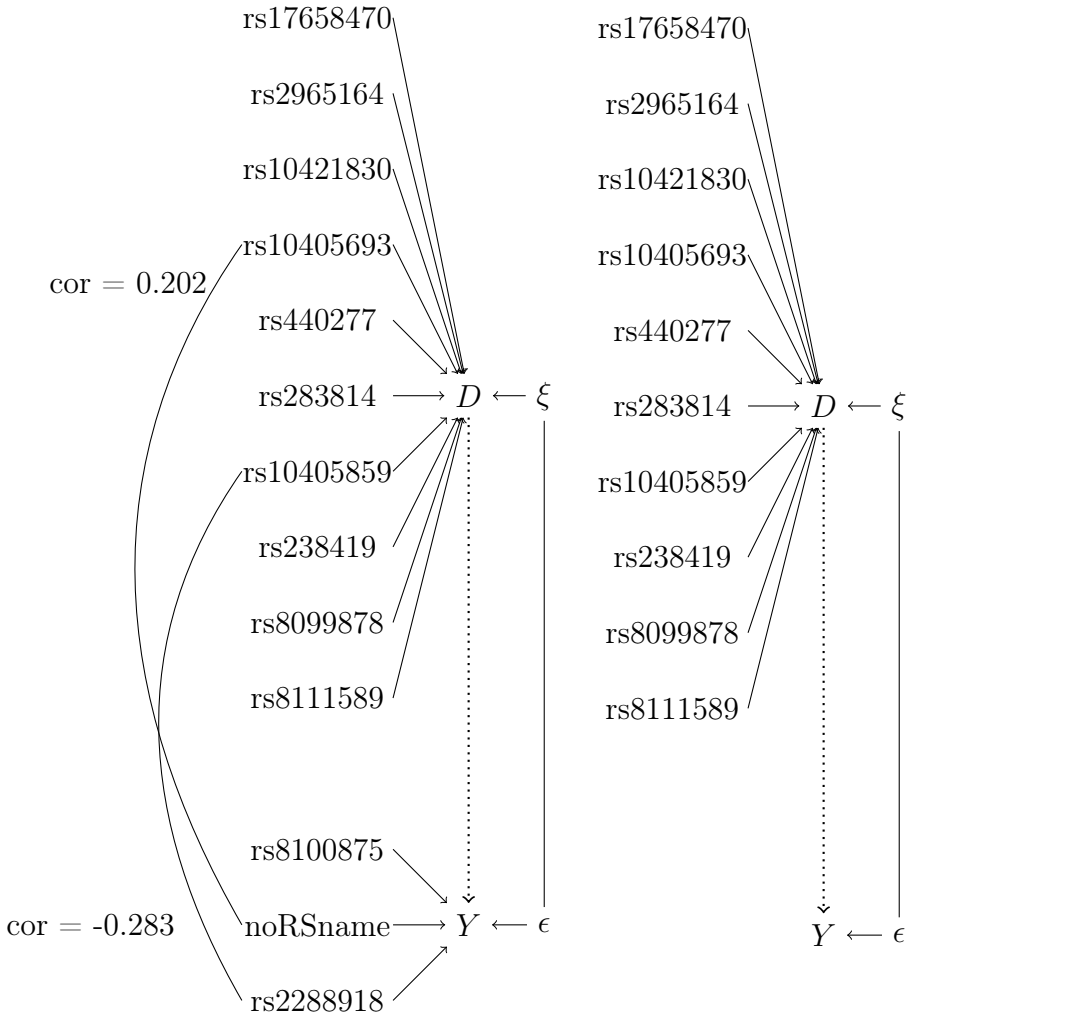
The setups of generating simulated data are shown in Figure C.3, where we only mark out $cor(rs10405693, noRSname) = 0.202$ and $cor(rs10405859, rs2288918) = -0.283$ because these two correlations were most significant. It is noted that besides

these two correlations, there existed correlations (due to linkage disequilibrium) between the 10 SNPs relevant to D and 3 invalid SNPs having direct effects on Y as shown in Table C.2. With Setup 1, for each β we did the simulation 1000 times and calculated the proportion of p-values less than 0.05. We compared three methods: the naive-2SLS, oracle-2SLS and 2ScML. For all three methods, we used the 10 SNPs to get $\hat{\boldsymbol{D}}$ in stage 1. For 2ScML, in stage 2 we included all 253 SNPs, set $\tau_2 = 1 \times 10^{-5}$, and used 5-fold cross-validation to choose the best K_2 from 0 to 6. For the naive-2SLS, in stage 2 we performed linear regression of Y on $\hat{\boldsymbol{D}}$; for the oracle-2SLS, in stage 2 we performed linear regression of Y on $\hat{\boldsymbol{D}}$, $rs8100875$, $noRSname$ and $rs2288918$. Figure C.4 shows the empirical Type-I Error rates when $\beta = 0$, and power when $\beta \neq 0$ from 1000 simulations.

In Figure C.4, the dashed horizontal black line is $y = 0.05$. When $\beta = 0$, it was the null case. We can see that at this setting, both 2ScML and oracle-2SLS could control the Type I Error rate at 0.05, but the naive-2SLS had an inflated Type I Error rate around 45%. As the absolute value of β increased, the power of 2ScML and oracle-2SLS increased. We observe that as β increased from 0 to 1, the power of the naive-2SLS decreased, which was due to the reason as we explained for Figure C.2; here a positive value of β in the range of 0 to 1 would diminish the total effect of the SNPs on Y and thus decrease the power of the naive-2SLS.

In Setup 2, the naive-TWAS was identical to the oracle-2SLS. For each of β, we did the simulation 1000 times and calculated the proportion of p-values less than 0.05. We compared two methods: oracle-2SLS and 2ScML. For both methods, we used the 10 SNPs to get $\hat{\boldsymbol{D}}$ in stage 1. For 2ScML, in stage 2 we included all 253 SNPs, set $\tau_2 = 1 \times 10^{-5}$, and used 5-fold cross-validation to choose the best K_2 from 0 to 6. For oracle-2SLS, in stage 2 we performed linear regression of Y on $\hat{\boldsymbol{D}}$ only. Figure C.5 shows the Type-I Error rate when $\beta = 0$, and power when $\beta \neq 0$ from the simulations.

Figure C.3: Generating simulated data for gene GEMIN7 in two setups.



(a) Setup 1 with 3 invalid IVs.

(b) Setup 2 with all IVs being valid.

Figure C.4: Simulation results for gene GEMIN7 with 3 IVs being invalid.

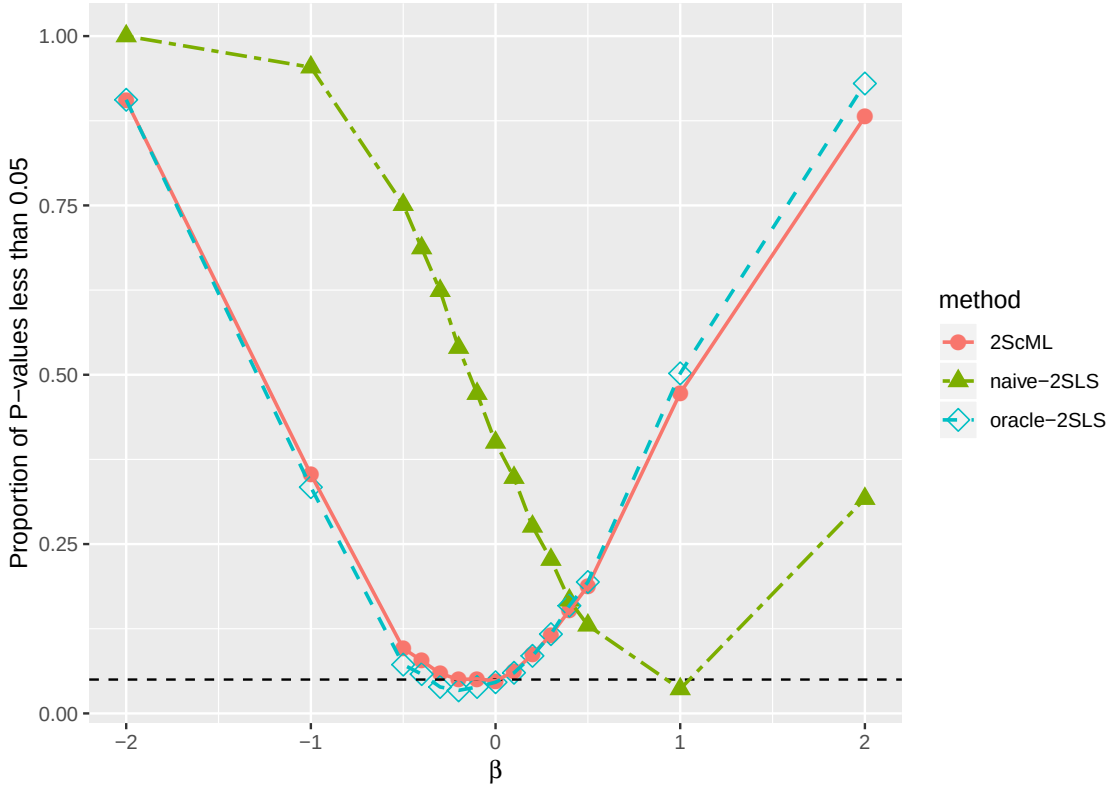
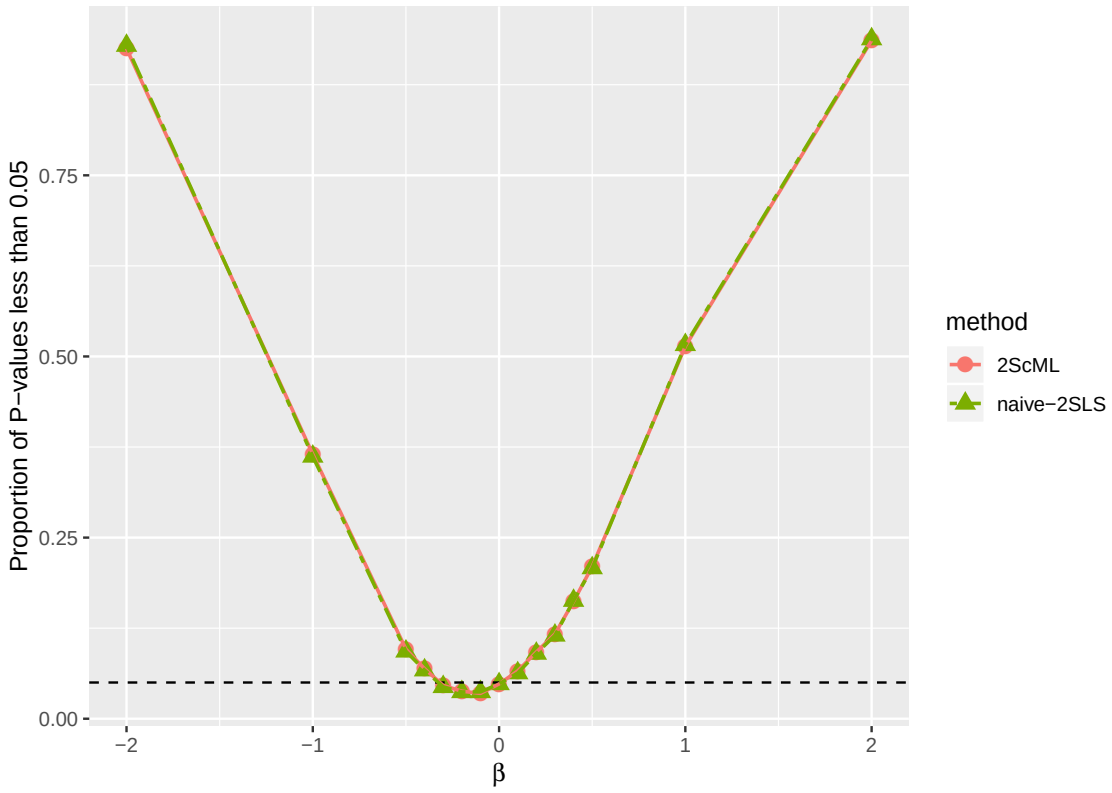


Figure C.5: Simulation results for gene GEMIN7 with all IVs being valid.



In Figure C.5, the dashed horizontal black line is $y = 0.05$. When $\beta = 0$ it was the null case. We can see in Setup 2, both 2ScML and oracle-2SLS could control the Type I Error rate at 0.05, and they performed similarly with almost the same power.

Appendix D

Real Data Example: More Results

Here we show the details of the 32 significant genes identified by TWAS and 2ScML in Table D.1.

Next we show the full literature search results for these 32 genes.

LDLRAP1 on Chromosome 1: LDLRAP1 is a liver-enriched gene that plays a critical role in facilitating the removal of LDL from the circulatory system [5, 6].

PIGV on Chromosome 1: From GWAS Catalog, PIGV has been reported as associated with LDL and HDL [7]. PIGV is a mannosyltransferase that plays a role in multiple cellular processes, including protein sorting and signal transduction [8]. And Glycosylphosphatidylinositol (GPI) is a complex glycolipid that anchors many proteins to the cell surface. The biosynthetic pathway of GPI is mediated by sequential addition of sugars and other components to phosphatidylinositol. PIGV adds the second mannose to the GPI core[9].

DOCK7 on Chromosome 1: From GWAS Catalog, genetic variants mapped to

Table D.1: The 32 significant genes associated with LDL identified by TWAS or/and 2ScML with their literature search support scores (Score) and corresponding references (Refs). The p-values less than the significance cut-off 0.05/4580 are marked red.

Gene	Chr	p	Best K_2	p_{TWAS}	p_{2ScML}	Score	Refs
LDLRAP1	1	13	1	2.35e-05	1.15e-07	5	[5], [6]
PIGV	1	5	1	6.58e-04	2.34e-06	3	[7], [8], [9]
DOCK7	1	13	0	7.64e-17	7.64e-17	3	[10]
PSRC1	1	19	5	8.40e-99	2.55e-75	3	[10]
PSMA5	1	13	6	1.12e-07	1.02e-03	2	[11], [12]
GNAI3	1	26	8	9.93e-06	7.11e-01	4	[13]
GSTM4	1	42	4	3.82e-06	9.65e-02	1	[14]
CCDC93	2	3	1	4.46e-03	3.39e-07	5	[15], [16], [17], [18]
MKRN2	3	12	0	4.92e-08	4.92e-08	2	[19]
RAF1	3	7	1	3.00e-04	8.63e-07	2	[20], [12]
PARP9	3	9	0	1.02e-06	1.02e-06	2	[16]
AIF1	6	15	1	1.20e-03	4.10e-06	2	[21], [22], [23]
DDAH2	6	13	0	3.87e-06	3.87e-06	4	[24], [25]
NOTCH4	6	18	1	1.13e-03	1.49e-06	2	[26], [21], [22]
TAP2	6	23	1	1.43e-03	4.37e-07	2	[27], [28], [29]
DDX56	7	9	2	6.46e-06	6.70e-02	3	[30], [19]
TMED4	7	11	1	7.06e-09	6.03e-04	0	NA
PARP10	8	4	0	1.35e-10	1.35e-10	3	[30]
GRINA	8	5	0	8.26e-11	8.26e-11	0	NA
FADS1	11	15	1	5.87e-14	2.85e-01	3	[26]
SH2B3	12	11	0	3.07e-10	3.07e-10	3	[31], [32], [33], [30], [34]
OASL	12	25	1	4.37e-08	1.22e-12	3	[35], [36], [37], [35]
HP	16	14	1	3.31e-03	3.84e-08	4	[16], [38], [39]
DHX38	16	10	3	4.33e-04	5.19e-09	3	[22], [40], [41], [42], [43]
TBKBP1	17	3	0	4.86e-06	4.86e-06	2	[16], [44]
KRI1	19	15	4	2.60e-01	2.49e-13	3	[30]
CARM1	19	9	2	4.02e-06	1.24e-01	4	[30], [45], [46], [47], [48]
SMARCA4	19	3	1	1.01e-25	3.62e-08	4	[16], [32], [49], [50]
LPAR2	19	6	1	1.80e-01	3.73e-10	4	[51]
PVRL2	19	11	4	1.24e-11	1.38e-05	3	[52], [53], [54], [55]
TOMM40	19	28	10	8.83e-36	3.48e-15	3	[56], [57], [58]
MAFB	20	4	1	5.76e-06	7.03e-01	3	[26], [59], [16]

this gene are shown associated with LDL, TG, TC [10].

PSRC1 on Chromosome 1: From GWAS Catalog, genetic variants mapped to this gene are shown associated with LDL, TC [10].

PSMA5 on Chromosome 1: From GWAS Catalog, PSMA5 has been reported as associated with Intelligence [11], and Body Mass Index [12].

GNAI3 on Chromosome 1: GNAI3 participates in the development of NAFLD in both cellular and mouse models [13].

GSTM4 on Chromosome 1: It is reported that a polymorphism in the GSTM4 gene implicated lung cancer risk [14].

CCDC93 on Chromosome 2: From GWAS Catalog, CCDC93 has been reported as associated with Venous thromboembolism [15], Triglycerides [16], Cognitive performance [17]. Also, [18] provides evidence that a common variant in CCDC93, encoding a protein involved in recycling of the LDLR, is associated with lower LDL-c levels, lower risk of myocardial infarction and cardiovascular mortality.

MKRN2 on Chromosome 3: From [19], MKRN2 is associated with systolic blood pressure, red cell distribution width, and eosinophil counts.

RAF1 on Chromosome 3: From GWAS Catalog, RAF1 has been reported as associated with cardiac hypertrophy [20], waist-to-hip ratio adjusted for BMI [12].

PARP9 on Chromosome 3: From GWAS Catalog, PARP9 has been re-

ported as associated with total cholesterol levels [16]

AIF1 on Chromosome 6: From GWAS Catalog, AIF1 has been reported as associated with metabolite levels [21], Blood protein levels [22]. And SNP rs2844479 in AIF1 contributes to obesity risk in the Greek population [23].

DDAH2 on Chromosome 6: Asymmetric dimethylarginine (ADMA), present in human serum, is an endogenous inhibitor of nitric oxide synthase and contributes to vascular disease. Genetic variation in DDAH2 gene is significantly associated with serum ADMA levels in participants with type 2 diabetes [24]. And hypermethylation in DDAH2 promoter is positively correlated to the dysfunction of endothelial progenitor cells (EPCs) in CAD patients [25].

NOTCH4 on Chromosome 6: From GWAS Catalog, NOTCH4 has been reported as associated with triglycerides [26], metabolite levels [21], blood protein levels [22].

TAP2 on Chromosome 6: From GWAS Catalog, TAP2 has been reported as associated with diastolic blood pressure [27], type 1 diabetes and autoimmune thyroid diseases [28], serum complement C3 and C4 levels [29].

DDX56 on Chromosome 7: From GWAS Catalog, DDX56 has been reported as associated with LDL and TC [30], and cardiovascular disease [19].

TMED4 on Chromosome 7: Have not found any related study about this gene.

PARP10 on Chromosome 8: From GWAS Catalog, PARP10 has been reported as associated with LDL and TC [30].

GRINA on Chromosome 8: Have not found any related study about this gene.

FADS1 on Chromosome 11: From GWAS Catalog, FADS1 has been reported as associated with LDL and TC [26].

SH2B3 on Chromosome 12: From GWAS Catalog, SH2B3 has been reported as associated with blood pressure [31], coronary heart disease [32], blood metabolite levels [33], LDL, HDL and TC [30]. And it is involved in blood diseases, autoimmune disorders, and vascular disease [34].

OASL on Chromosome 12: From GWAS Catalog, OASL has been reported as associated with cardiovascular disease risk factors [35], type 2 diabetes [36], serum metabolite levels [37]. And OASL showed effects on gamma glutamyltransferase, LDL and C-reactive protein [35].

HP on Chromosome 16: From GWAS Catalog, HP has been reported as associated with LDL, HDL, TC [16]. And HP is linked to diabetic nephropathy [38], and incidence of coronary artery disease in type 1 diabetes [39].

DHX38 on Chromosome 16: From GWAS Catalog, DHX38 has been reported as associated with blood protein levels [22], interaction between LDL and sleep [40], coronary artery disease [41], serum metabolite levels [42]. And from [43], DHX38 is related to pig growth rate.

TBKBP1 on Chromosome 17: From GWAS Catalog, TBKBP1 has been reported as associated with HDL [16], body mass index [44].

KRI1 on Chromosome19: From GWAS Catalog, KRI1 has been reported as associated with LDL and TC [30].

CARM1 on Chromosome 19: From GWAS Catalog, CARM1 has been reported as associated with LDL and TC [30], C-reactive protein levels [45]. And C9orf72-CARM1 axis in the control of stress-induced lipid metabolism and implicates epigenetic dysregulation in relevant human diseases [46]. [47] shows that CARM1 promotes adipocyte differentiation by coactivating PPARγ-mediated transcription and thus might be important in energy balance. [48] shows that increased CARM1 expression in type 2 diabetes suggests that epigenetic mechanisms are altered in human diabetes.

SMARCA4 on Chromosome 19: From GWAS Catalog, SMARCA4 has been reported as associated with LDL and TC [16], coronary heart disease [32], peripheral artery disease [49]. And [50] suggests that SMARCA4 polymorphism may conducive to play a protective role against the hypertension risk.

LPAR2 on Chromosome 19: [51] supports LPAR2 as a potential effector gene in the fatty liver NCAN locus.

PVRL2 on Chromosome 19: From GWAS Catalog, AC011481.2, which is antisense to PVRL2, has been reported as associated with body fat percentage [52], type 2 diabetes [53], LDL and TG[54], body mass index [55].

TOMM40 on Chromosome 19: From GWAS Catalog, TOMM40 has been reported as associated with LDL, TG, TC, C-reactive protein [56], Alzheimer's disease [57], metabolic syndrome [58].

MAFB on Chromosome 20: From GWAS Catalog, MAFB has been reported as associated with LDL, TC [26], serum total protein level [59], triglycerides [16].

References for Appendix

[1] Guo, Z., Kang, H., Tony Cai, T., & Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80*(4), 793-815.

[2] Shen, X., Pan, W., Zhu, Y., & Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics, 65*(5), 807-832.

[3] Windmeijer, F., Farbmacher, H., Davies, N., & Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association, 114*(527), 1339-1350.

[4] Zhu, Y., Shen, X., & Pan, W. (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association, 115*(529), 217-230.

[5] Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., ... & Yin, Y. (2012). Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell research, 22*(1), 107-126.

[6] Garcia, C. K., Wilund, K., Arca, M., Zuliani, G., Fellin, R., Maioli, M., ... & Barnes, R. (2001). Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science, 292*(5520), 1394-1398.

[7] Spracklen, C. N., Chen, P., Kim, Y. J., Wang, X., Cai, H., Li, S., ... & Wu, J. Y. (2017). Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Human molecular genetics, 26*(9), 1770-1784.

[8] Krawitz, P. M., Schweiger, M. R., Rodelsperger, C., Marcelis, C., Kolsch, U., Meisel, C., ... & Isau, M. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nature genetics, 42*(10), 827-829.

[9] Kang, J. Y., Hong, Y., Ashida, H., Shishioh, N., Murakami, Y., Morita, Y. S., ... & Kinoshita, T. (2005). PIG-V involved in transferring the second mannose in glycosylphosphatidylinositol. *Journal of Biological Chemistry, 280*(10), 9489-9497.

[10] Aulchenko, Y. S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I. M., Pramstaller, P. P., ... & Martin, N. G. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics, 41*(1), 47.

[11] Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., ... & Grasby, K. L. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics, 50*(7), 912-919.

[12] Zhu, Z., Guo, Y., Shi, H., Liu, C. L., Panganiban, R. A., Chung, W., ... & Camargo Jr, C. A. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *Journal of Allergy and Clinical Immunology, 145*(2), 537-549.

[13] Zhu, H., Ge, K., Lu, J., & Jia, C. (2019). Downregulation of GNAI3 Promotes the Pathogenesis of Methionine/Choline-Deficient Diet-Induced Nonalcoholic Fatty Liver Disease. *Gut and liver*.

[14] Liloglou, T., Walters, M., Maloney, P., Youngson, J., & Field, J. K. (2002). A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer. *Lung cancer, 37*(2), 143-146.

[15] Deguchi, H., Shukla, M., Hayat, M., Torkamani, A., Elias, D. J., & Griffin, J. H. (2020). Novel exomic rare variants associated with venous thrombosis. *British journal of haematology*, 190(5), 783-786.

[16] Klarin, D., Damrauer, S. M., Cho, K., Sun, Y. V., Teslovich, T. M., Honerlaw, J., ... & Chaffin, M. (2018). Genetics of blood lipids among 300,000 multi-ethnic participants of the Million Veteran Program. *Nature genetics, 50*(11), 1514-1523.

[17] Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... & Fontana, M. A. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics, 50*(8), 1112-1121.

[18] Rimbert, A., Dalila, N., Wolters, J. C., Huijkman, N., Smit, M., Kloosterhuis, N., ... & Biobank-Based Integrative Omics Studies Consortium. (2020). A common variant in CCDC93 protects against myocardial infarction and cardiovascular mortality by regulating endosomal trafficking of low-density lipoprotein receptor. *European heart journal, 41*(9), 1040-1053.

[19] Kichaev, G., Bhatia, G., Loh, P. R., Gazal, S., Burch, K., Freund, M. K., ... & Price, A. L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics, 104*(1), 65-75.

[20] Parsa, A., Chang, Y. P. C., Kelly, R. J., Corretti, M. C., Ryan, K. A., Robinson, S. W., ... & Liggett, S. B. (2011). Hypertrophy-associated polymorphisms ascertained in a founder cohort applied to heart failure risk and mortality. *Clinical and translational science, 4*(1), 17-23.

[21] Chai, J. F., Raichur, S., Khor, I. W., Torta, F., Chew, W. S., Herr, D. R., ... & Tai, E. S. (2020). Associations with metabolites in Chinese suggest new metabolic roles in Alzheimer's and Parkinson's diseases. *Human Molecular Genetics, 29*(2), 189-201.

[22] Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., ... & Oliver-Williams, C. (2018). Genomic atlas of the human plasma proteome. *Nature, 558*(7708), 73-79.

[23] Rouskas, K., Kouvatsi, A., Paletas, K., Papazoglou, D., Tsapas, A., Lobbens, S., ... & Meyre, D. (2012). Common variants in FTO, MC4R, TMEM18, PRL, AIF1, and PCSK1 show evidence of association with adult obesity in the Greek population. *Obesity, 20*(2), 389-395.

[24] Abhary, S., Burdon, K. P., Kuot, A., Javadiyan, S., Whiting, M. J., Kasmeridis, N., ... & Craig, J. E. (2010). Sequence variation in DDAH1 and DDAH2 genes is strongly and additively associated with serum ADMA concentrations in individuals with type 2 diabetes. *PLoS One, 5*(3).

[25] Niu, P. P., Cao, Y., Gong, T., Guo, J. H., Zhang, B. K., & Jia, S. J. (2014). Hypermethylation of DDAH2 promoter contributes to the dysfunction of endothelial progenitor cells in coronary artery disease patients. *Journal of translational medicine, 12*(1), 170.

[26] Surakka, I., Horikoshi, M., Magi, R., Sarin, A. P., Mahajan, A., Lagou, V., ... & Kettunen, J. (2015). The impact of low-frequency and rare variants on lipid levels. *Nature genetics, 47*(6), 589.

[27] Warren, H. R., Evangelou, E., Cabrera, C. P., Gao, H., Ren, M., Mifsud, B., ... & Kraja, A. T. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature genetics, 49*(3), 403.

[28] Tomer, Y., Dolan, L. M., Kahaly, G., Divers, J., D'Agostino Jr, R. B., Imperatore, G., ... & Hasham, A. (2015). Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes. *Journal of autoimmunity, 60*, 32-39.

[29] Yang, X., Sun, J., Gao, Y., Tan, A., Zhang, H., Hu, Y., ... & Kim, S. T. (2012). Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. *PLoS genetics, 8*(9).

[30] Hoffmann, T. J., Theusch, E., Haldar, T., Ranatunga, D. K., Jorgenson, E., Medina, M. W., ... & Iribarren, C. (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nature genetics, 50*(3), 401-413.

[31] Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., ... & Aulchenko, Y. (2009). Genome-wide association study of blood pressure and hypertension. *Nature genetics, 41*(6), 677.

[32] Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., ... & Absher, D. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics, 43*(4), 333-338.

[33] Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., ... & Walter, K. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics, 46*(6), 543.

[34] Auburger, G., Gispert, S., Lahut, S., Omur, O., Damrath, E., Heck, M., & Basak, N. (2014). 12q24 locus association with type 1 diabetes: SH2B3 or ATXN2? *World journal of diabetes, 5*(3), 316.

[35] Middelberg, R. P., Ferreira, M. A., Henders, A. K., Heath, A. C., Madden, P. A., Montgomery, G. W., ... & Whitfield, J. B. (2011). Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits. *BMC medical genetics, 12*(1), 123.

[36] Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., ... & McCulloch, L. J. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics, 42*(7), 579.

[37] Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V., ... & Gieger, C. (2012). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS genetics, 8*(10).

[38] Asleh, R., & Levy, A. P. (2005). In vivo and in vitro studies establishing haptoglobin as a major susceptibility gene for diabetic vascular disease. *Vascular health and risk management, 1*(1), 19.

[39] Sadrzadeh, S. H., & Bozorgmehr, J. (2004). Haptoglobin phenotypes in health and disorders. *Pathology Patterns Reviews, 121*(suppl_1), S97-S104.

[40] Noordam, R., Bos, M. M., Wang, H., Winkler, T. W., Bentley, A. R., Kilpelainen, T. O., ... & Manning, A. (2019). Multi-ancestry sleep-by-SNP interaction anal-

ysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nature communications, 10*(1), 1-13.

[41] van der Harst, P., & Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research, 122*(3), 433-443.

[42] Gallois, A., Mefford, J., Ko, A., Vaysse, A., Julienne, H., Ala-Korpela, M., ... & Aschard, H. (2019). A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nature communications, 10*(1), 1-13.

[43] Li, X., Kim, S. W., Do, K. T., Ha, Y. K., Lee, Y. M., Yoon, S. H., ... & Kim, K. S. (2011). Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Molecular biology reports, 38*(6), 3805-3820.

[44] Hoffmann, T. J., Choquet, H., Yin, J., Banda, Y., Kvale, M. N., Glymour, M., ... & Jorgenson, E. (2018). A large multiethnic genome-wide association study of adult body mass index identifies novel loci. *Genetics, 210*(2), 499-515.

[45] Han, X., Ong, J. S., An, J., Hewitt, A. W., Gharahkhani, P., & MacGregor, S. (2020). Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration. *European Journal of Epidemiology*, 1-8.

[46] Liu, Y., Wang, T., Ji, Y. J., Johnson, K., Liu, H., Johnson, K., ... & Wang, J. (2018). A C9orf72-CARM1 axis regulates lipid metabolism under glucose starvation-induced nutrient stress. *Genes & development, 32*(21-22), 1380-1397.

[47] Yadav, N., Cheng, D., Richard, S., Morel, M., Iyer, V. R., Aldaz, C. M., & Bedford, M. T. (2008). CARM1 promotes adipocyte differentiation by coactivating PPARγ. *EMBO reports, 9*(2), 193-198.

[48] Porta, M., Amione, C., Barutta, F., Fornengo, P., Merlo, S., Gruden, G., ... & Beguinot, F. (2019). The co-activator-associated arginine methyltransferase 1 (CARM1) gene is overexpressed in type 2 diabetes. *Endocrine, 63*(2), 284-292.

[49] Klarin, D., Lynch, J., Aragam, K., Chaffin, M., Assimes, T. L., Huang, J., ... & Arya, S. (2019). Genome-wide association study of peripheral artery disease in the Million Veteran Program. *Nature medicine, 25*(8), 1274-1279.

[50] Ma, H., He, Y., Bai, M., Zhu, L., He, X., Wang, L., & Jin, T. (2019). The genetic polymorphisms of ZC3HC1 and SMARCA4 are associated with hypertension risk. *Molecular genetics & genomic medicine, 7*(11), e942.

[51] DiStefano, J. K., Kingsley, C., Wood, G. C., Chu, X., Argyropoulos, G., Still, C. D., ... & Gerhard, G. S. (2015). Genome-wide analysis of hepatic lipid content in extreme obesity. *Acta diabetologica, 52*(2), 373-382.

[52] Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., ... & Evans, D. M. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications, 7*(1), 1-15.

[53] Cook, J. P., & Morris, A. P. (2016). Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal of Human Genetics, 24*(8), 1175-1180.

[54] Bentley, A. R., Sung, Y. J., Brown, M. R., Winkler, T. W., Kraja, A. T., Ntalla, I., ... & Guo, X. (2019). Multi-ancestry genome-wide gene-smoking interaction

study of 387,272 individuals identifies new loci associated with serum lipids. *Nature genetics, 51*(4), 636-648.

[55] Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., ... & Yang, J. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human molecular genetics, 28*(1), 166-174.

[56] Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., ... & Belbin, G. M. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature, 570*(7762), 514-518.

[57] Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., ... & Jones, N. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature genetics, 41*(10), 1088.

[58] Kristiansson, K., Perola, M., Tikkanen, E., Kettunen, J., Surakka, I., Havulinna, A. S., ... & Eriksson, J. G. (2012). Genome-wide screen for metabolic syndrome susceptibility Loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circulation: Cardiovascular Genetics, 5*(2), 242-249.

[59] Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., ... & Kubo, M. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature genetics, 50*(3), 390-400.